UCSF
UC San Francisco Electronic Theses and Dissertations

Title
Protein folding with homologous sequences

Permalink
https://escholarship.org/uc/item/1cb148hd

Author
DeFay, Thomas Robert

Publication Date
1996

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

PROTEIN FOLDING WITH HOMOLOGOUS SEQUENCES

by

Thomas Robert DeFay

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biophysics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA

San Francisco

Date **University Librarian**

Degree Conferred: . . . . . .

To 'Ris, with Love

Acknowledgments

First I wish to thank the most important person in my life, my wife Marisa. From my first interview at UCSF, to my last days as a graduate student, she has been my support and my joy. I could not have accomplished what I have without her.

Next, are of course my Mom and Dad. Unlike many people I have met, I have great parents. They taught me to be independent and follow my dreams. I know they love me as I love them, and I hope to follow in their footsteps when I am a parent.

Next I wish to thank my advisor Fred Cohen. The most fitting way I can think of to explain what Fred has meant to me as a scientist is through an analogy. The best player on my soccer team was a guy named Uli. During a game, again and again, the ball would appear out of nowhere, in front of the goal mouth, directly in front of my foot. It seemed to hang there, waiting for me. So I pounded it into the back of the net. At the end of the game, sometimes after scoring five goals, I thought I was the best soccer player ever. Upon further reflection, I realized that Uli had made every one of those passes. He was never flashy about it, didn't pick you up and carry you around the field afterwards, but I couldn't have done it without him. Fred has given me the perfect pass many times. For me, his lab has been the ideal scientific environment. This has included a myriad of specific oppotunities, but perhaps more importantly, Fred has gone out of his way to attract helpful, interesting people.

The most helpful of these was a scientist named Scott Presnell. Scott was always willing to help others, a trait I took advantage of on embarrasingly

numerous occasions. The karmic debt I accrued learning C programming, UNIX, and protein folding, will take years of assistance to green graduate students to pay off. Scott was one of the main reasons I joined the Cohen group, and is in large part responsible for my skills as a computational scientist.

Scott was not alone, so many people in the Cohen lab have provided their support and friendship. In particular, I would like to thank Yang for her friendship and her hard work on Factor VIII. Ginger, for being my lab mother, an appelation that bugs her, but we all really appreciate it. To Alexis, for his enthusiasm, and completion of the soap bubble that took a year off of my time in graduate school. To Chris, who's scientific approach I have tried to emulate, although I do not plan on going to law school. To John, for being a friend all of these years, and covering for Scott when I had still more questions. To Nathan, with whom I spent many an hour playing GO. I will miss your philosophy, you have been a good friend. So many others, Mike, Dietlind, Paul, Xiauwu, Olivier, Phillippe, Roland, Steve, its been great.

Protein Folding with Homologous Sequences

by

Thomas R. DeFay

Abstract

The protein folding problem has been one of the most intractable problems facing science for almost 40 years. The problem is to predict the three-dimensional structure of a protein from its amino acid sequence. Early on, it was hoped that a simple pattern relating the amino acids would help solve this problem, much as the structure of DNA was solved. When this proved unsuccessful, efforts turned toward developing energy functions accurate enough to identify the native structure. In forty years this problem still has not yielded.

Fortunately, a homologous family of sequences all fold to a similar three-dimensional structure. This fact can be exploited to increase the accuracy of structure predictions. The most straightforward way to use a homologous sequence is if that sequence already has an experimentally determined structure. In this case, the structure can be used as a template upon which to build a new structure, as we have done.

We have also proposed a new method for predicting structure based upon an old technique, threading. When threading, the goal is to match a sequence with one of a set of known folds in a protein database. We have devised a new method which "threads" using the information from a set of homologous sequences.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Background

How does the sequence of a polypeptide lead to its three dimensional structure? In the case of DNA, knowledge of its three-dimensional structure readily led to an understanding of the binding and interaction of base-pairs to form the structure of DNA, namely the double helix. In the case of proteins, the connection between sequence and structure is less clear. John Kendrew and co-workers remarking on the structure of myoglobin said "Perhaps the most remarkable features of the molecule are its complexity and lack of symmetry (Kendrew *et al.*, 1958)." This was unfortunate due to the importance of the structure of proteins. Their structure is responsible for their function, and they perform most of the functions in the human body.

Hope emerged for finding the connection between sequence and structure for proteins, when in the early 1970's, Anfinsen established that the information required to fold ribonuclease was contained solely in its polypeptide chain (Anfinsen *et al.*, 1961). This implies that the native structure of ribonuclease (and it turns out, most globular proteins) is at a thermodynamic energy minimum.

Given that the native structure of a protein is at its free energy minimum, the protein folding problem can be solved by generating an accurate free energy function, and searching the conformations of the protein chain for the one that is of lowest energy. At present, both of these steps have proven to be intractable.

Efforts have continued to find local patterns in the amino acid sequence that have structurally predictive value. Some headway has been made in the area of secondary structure prediction, but accuracy for this simplified problem is still near 63% (Garnier & Levin, 1991).

Another technique has also emerged, called "threading" which involves matching a sequence onto each of a set of structures in a database, and choosing the best match. The effectiveness of this technique is based on the fact that unrelated protein sequences often fold to similar overall structures. At first, this technique looked quite promising, but recent results from a conference at Asilomar suggest that our ability to thread accurately is still quite low (Lemer *et al.*, 1995).

Luckily, we are not limited to the information in just one amino acid sequence. Sequences from the same family of sequences fold to very similar three-dimensional structures. This information has not been exploited to generate new and accurate energy functions, but has benefited approaches based upon statistical patterns. Thus, secondary structure prediction accuracy improves to 72% when homologous sequence information is exploited (Rost & Sander, 1994). This increased accuracy has led in some cases to accurate structure predictions based on assembling the secondary structure units (Benner *et al.*, 1994; Crawford *et al.*, 1987).

In addition, in some cases the three dimensional structure of a sequence that is homologous to the sequence of a new protein has been solved. In this case, a technique called homology modeling can be used to predict the structure of the new sequence.

UCSF LIBRARY

Thesis

The second chapter in this thesis focuses on the overall problem of protein modeling, touching on approaches such as secondary structure prediction, energy functions and homology modeling. It serves as a basis for understanding the following chapters.

The third chapter is devoted to a proven method for exploiting homologous sequences--homology modeling. In this case, a homology model of protein factor VIII was generated. Defects in this protein may lead to hemophilia. Our homology model allowed us to explain much of the experimental evidence known about this protein, and its link to hemophilia A.

The fourth chapter in this thesis is an evaluation of a blind prediction contest held in Asilomar California. In this contest, investigators attempted to predict using *de novo* methods (basically, not homology modeling or threading methods) the structure of several proteins. We found that the most important ingredient for success was the inclusion of information from a set of homologous sequences, especially for secondary structure prediction.

One area of protein folding which has not yet benefited from the use of homologous sequences is threading. Some studies have matched predicted secondary structure elements generated using multiple sequence alignments with secondary structure elements from known folds. In the final chapter of this thesis, we present the first study where homologous sequences have been used to carry out explicit threadings to find the crystal structure that best matches the sequence in question.

References

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**, 1309-1313.

Benner, S. A., Badcoe, I., Cohen, M. A. & Gerloff, D. L. (1994). Bona Fide Prediction of Aspects of Protein Conformation. *J. Mol. Biol.* **235**, 926-958.

Crawford, I. P., Niermann, T. & Kirschner, K. (1987). Prediction of Secondary Structure by Evolutionary Comparison: Application to the alpha Subunit of Tryptophan Synthase. *Proteins* **2**, 118-129.

Garnier, J. & Levin, J. M. (1991). The protein structure code: what is its present status? *CABIOS* **7**, 133-142.

Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R. G., Wyckoff, H. & Phillips, D. C. (1958). A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature* **181**, 662-666.

Lemer, C. M. R., Rooman, M. J. & Wodak, S. J. (1995). Protein Structure Prediction By Threading Methods: Evaluation of Current Techniques. *Proteins* **23**, 337-355.

Rost, B. & Sander, C. (1994). Combining Evolutionary Information and Neural Networks to Predict Protein Secondary Structure. *Proteins* **19**, 55-72.

UCSF LIBRARY

Chapter 2

Protein Modeling

UCSF LIBRARY

This chapter has been accepted for publication by the Encyclopedia of Molecular Biology and Molecular Medicine, VCH Publishers, Weinheim, Germany for 1996

Introduction

Protein Modeling is a collection of computational methods for the description, analysis and prediction of protein structures and the interaction of proteins with other molecules. Protein structures can provide insight into enzyme mechanisms and are increasingly useful for the design and optimization of novel pharmaceuticals. A major goal of protein modeling is the "Protein Folding Problem," the ability to accurately predict the structure of a protein from its sequence. The protein folding problem is still considered intractable for most proteins when attempted without additional information (*de novo*). Headway is being made however, with restricted systems such as all helical proteins. Another approach, modeling by homology, takes advantage of an experimentally determined structure. This structure is used as a template for the structure of a protein with a similar (homologous) sequence.

Most representations of protein structures are static but, in solution, proteins are quite flexible and are constantly changing form. The static structure is an average of the most frequently observed positions of the atoms that constitute a protein. Increasingly dynamic models have been developed to follow the conformational plasticity of a protein over time.

USF LIBRARY

Importance of Protein Structure to Protein Function/Inhibitor Design

Origin of Structures: X-ray and NMR

Most protein structures have been determined by the technique of X-ray crystallography. For this method to succeed, suitable conditions must be found to induce the protein into a highly ordered crystalline array capable of coherently scattering X-rays. The intensities of the diffracted X-rays are measured but the phase information is lost. When visual light is diffracted off of an object, the eye refocuses the light, combining intensities and phase information into an image. Lenses are not available that can refocus X-rays, and detectors cannot measure the phase of light. The phase information can be deduced by a variety of methods, including the introduction of heavy atoms into the crystal (multiple isomorphous replacement and multiple anomalous dispersion) or by relation to previously solved proteins of similar structure (molecular replacement). The coordinates of many structures are stored in the Brookhaven Protein Data Bank (PDB).

More recently, multi-dimensional Nuclear Magnetic Resonance (NMR) methods have been used to determine the structures of many proteins with an accuracy comparable to X-ray crystallography. NMR structures are determined in solution so the problem of obtaining crystals is eliminated. Technical limitations make it difficult to determine the structure of proteins larger than 20kD by NMR methods.

Biological Lessons Learned

The structures produced by X-ray Crystallography and NMR have been used to enhance our understanding of biological processes. For instance, the structure of myosin is helping to resolve the mechanics of muscle contraction. Myosin is thought to move along an actin filament in a series of

USF LIBRARY

steps. The structure of myosin is being used to discover the length of each step and the method of attachment of myosin to actin. Crystal structures have helped resolve the specific mechanism of enzyme catalysis in several cases. An enzyme catalyzes a chemical reaction in a cleft on its surface (the active site). Several different mutants of the enzyme can be made to determine which amino acid residues form the active site. When this information is combined with the protein structure, the mechanism of enzyme action may be deduced. Knowledge of the enzyme mechanism has been sufficiently detailed to change the substrate the enzyme performs its function on.

Recent applications to drug design

Drugs have been designed to bind to or alter the active site of an enzyme to block the reaction that occurs on its surface. If this is done to a protein important to the replication of a virus, the virus may die out. Unfortunately, the success of designed drugs has been limited by the complexities of the binding interactions. More headway has been made by identifying the basic shape of an active site and searching a large database of molecules to find a subset that geometrically match. These candidates for binding are then biologically assayed. Analysis of the structure of a biologically active compound complexed with a protein can lead to improvements in the compound. Several rounds of compound alteration and structural analysis may produce a usable drug.

Relationship of Sequence → Structure

The Thermodynamic Hypothesis

Most of protein modeling is based on the hypothesis that the final structure of a protein is uniquely determined by its amino acid sequence. This means that the final folded state of the protein is the thermodynamic minimum state for the protein under standard physiological conditions. A series of experiments by Anfinsen and coworkers demonstrated that a denatured protein (ribonuclease) will spontaneously refold in solution to adopt its active structure (**figure II.1**). Ribonuclease will refold even if the native disulfide pairings have been scrambled. This has been taken as evidence that proteins fold to reach their thermodynamically optimum state.

Relevance of Chaperonins

Molecular chaperones assist the folding of some proteins. It is thought that they catalyze the transition of a protein from its unfolded to folded state by enveloping a folding protein. This process may lower the activation barrier to the conformation search for the minimum energy state (**figure II.2**), and may prevent folding intermediates from aggregating into non-productive forms.

Kinetic Problems

Catalysts for the protein folding process may take other forms. α-lytic protease requires an N-terminal Pro-segment to fold correctly. Without it, the protein folds to a non-native like structure. Under normal folding conditions the pro region is autocatalytically removed after the protein folds. The pro region reduces the activation energy necessary to fold the protein (**figure II.2**). Without the pro-region the protease sequence cannot overcome

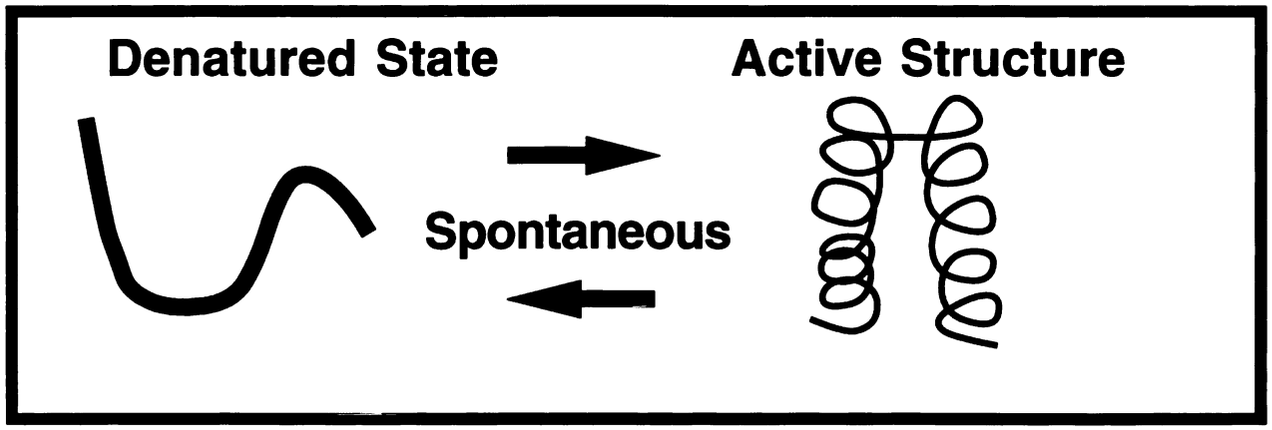


Figure II.1. Denaturation and renaturation. The protein changes to and from a random chain and a compact structure upon change of environmental conditions (such as solvent or temperature).

UCSF LIBRARY

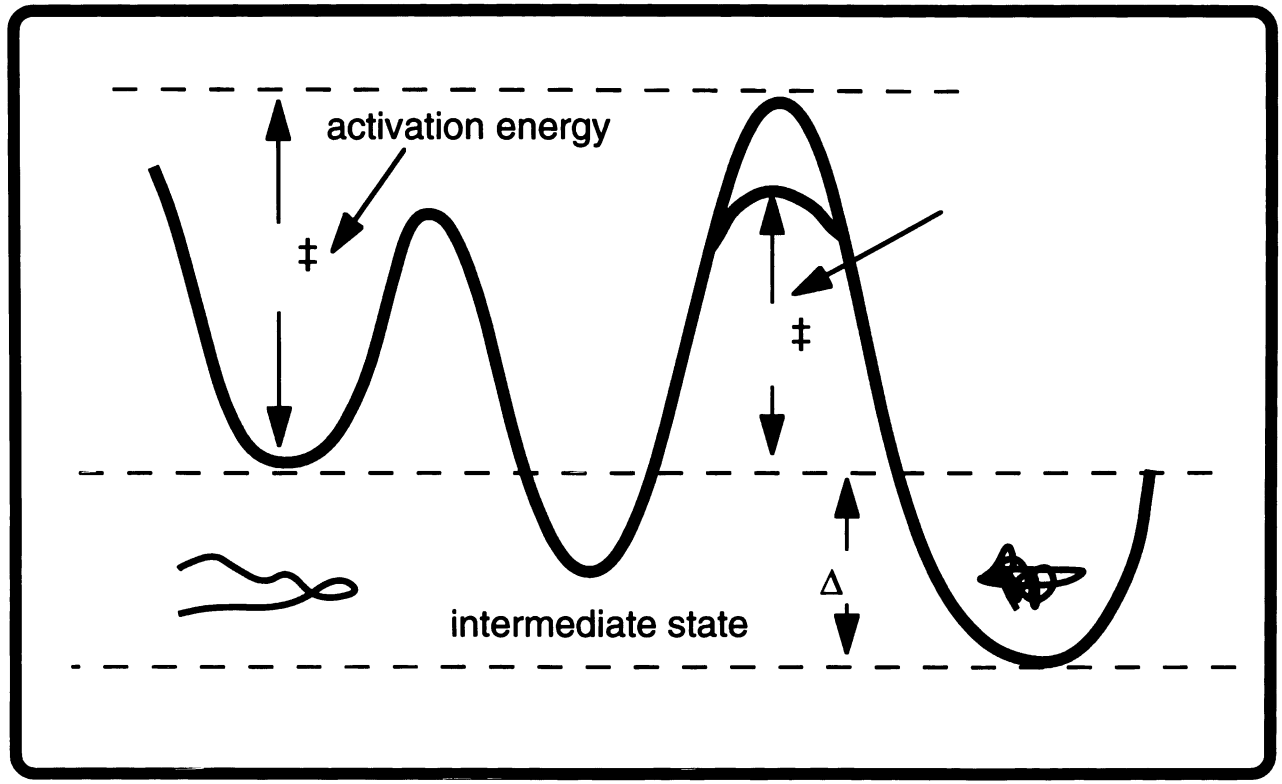


Figure II.2. Hypothetical activation energy profile for protein folding with a pro-region acting as a catalyst. DG‡ is the activation energy of folding for the protein with and without the presence of a catalyst. DG is the free energy of folding the protein. The activation energy of the folding process is reduced by the catalyst thus increasing the rate of the reaction. Note that the free energy of folding remains constant.

the ~126 kJ (30 kcal) activation barrier to folding on a biologically relevant time scale (< Hours).

Alternate Low Energy States

Proteins may have a state lower in energy than the folded state that is not kinetically accessible. However, this low energy state would have no evolutionary pressure to remain lowest in energy. Random mutations in thesequence would be much more likely to destabilize an unused state than the biologically important folded state. Over time, the low energy state would be destabilized causing the folded state to be lowest in energy. The fact that a protein sequence folds to one low energy structure does not imply that other protein sequences do not fold to the same low energy structure. Many different sequences may fold to the same structure. The relative stability of the structure may be quite different for different sequences, but the folded structure is the lowest energy state available for each sequence.

USF LIBRARY

Overview of Modeling Proteins

De novo vs. By Homology

Two general approaches to protein modeling are used: the *de novo* and Homology based methods. Homology modeling requires knowledge of the structure of a sequence that is recognizably similar to the desired protein. The known structure is used as a template upon which one engrafts the new sequence. For sequences that share identical residues at more than 30% of their aligned positions, model built structures can be quite accurate and have been used to design novel pharmaceuticals. Pharmaceuticals are designed to interact with the active site. The active site of a protein varies less than other regions, allowing this method to be possible. *De novo* modeling does not require the initial protein structure, instead a model is constructed from an analysis of the sequence in an attempt to produce a structure that is optimally suited to that sequence. Although *de novo* methods are unlikely to approach the accuracy of homology based strategies they are applicable in principle to a broader range of problems.

Static vs. Dynamic Structures

Rigid models of protein structures capture only one facet of the conformational properties of these macromolecules. A catalytically important residue, when viewed in a static model, may appear to be inaccessible to the ligand molecule. However, this residue may be accessible during a significant part of the molecular trajectory calculated in a dynamic simulation. X-ray crystallography determines an average set of atomic positions best represented by a static structure with a thermal motion or B-factor associated with mobile atoms. Intramolecular distance constraints derived by NMR spectroscopy are used to determine a family of structures

USF LIBRARY

consistent with the experimental data. The molecular envelope defined by the family of structures provides a more dynamic view of macromolecular conformation (**figure II.3**).

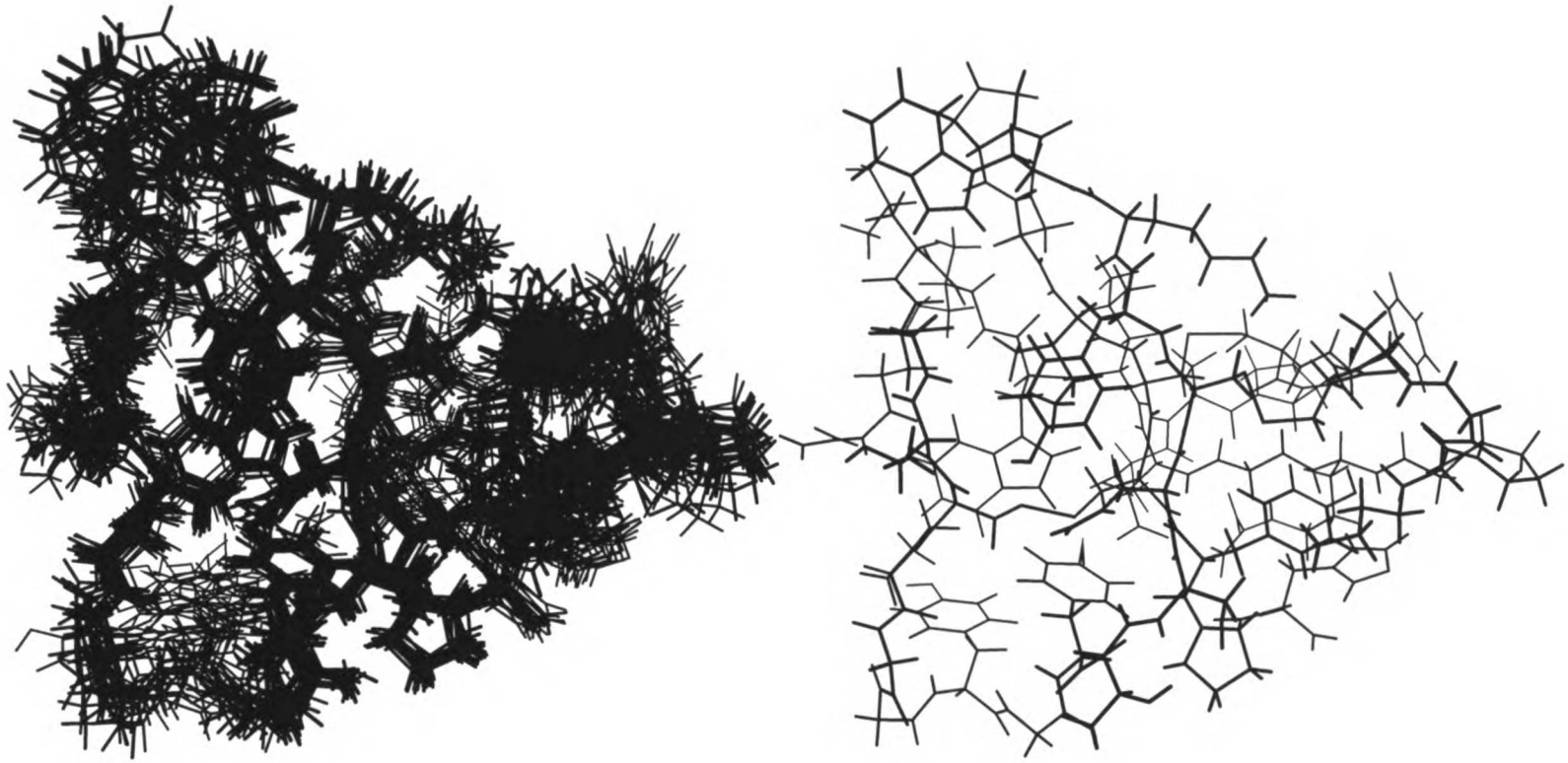


Figure II.3. BDS-1 envelope of twelve NMR structures, and a single averaged structure. The envelope reveals the flexibility of the protein chain exhibited by all proteins. The structures were taken from the Brookhaven Protein Data Base. They were solved by G.M. Clore.

UCSF MidasPlus

De Novo Methods

Secondary Structure Prediction

Most *de novo* modeling strategies attempt to identify the secondary structural elements of the protein from the protein sequence, and then assemble these structural elements into one or more plausible tertiary structures. These tertiary structures are evaluated for their compatibility with various experimental properties of the protein and/or theoretical properties of proteins in general.

Secondary structure prediction has evolved from early work by Schiffer and Edmunson who observed that helical sequences tended to segregate hydrophilic residues from their hydrophobic counterparts. A sequence is displayed as a wheel with consecutive spokes every 100 degrees (**figure II.4**). If the sequence forms a helix, the hydrophobic amino acids group together, as do the hydrophilic amino acids. The amino acids of sequences that form other secondary structures do not group in this manner.

Chou and Fasman compiled the frequency with which each amino acid appears in a specific secondary structural element. The structure of a sequence was then predicted using the probabilities generated from these frequencies. Garnier et al. improved on this by calculating the secondary structure preferences of each amino acid subject to modifications exerted by sequentially proximal residues. Recent recalibration of the Garnier technique has resulted in a secondary structure prediction algorithm that is ~65% accurate (Evaluation of secondary structure prediction accuracy is presented in **appendix 1**).

Neural networks, a computational tool from the machine learning community, have been designed to automatically identify patterns that

UCSF LIBRARY

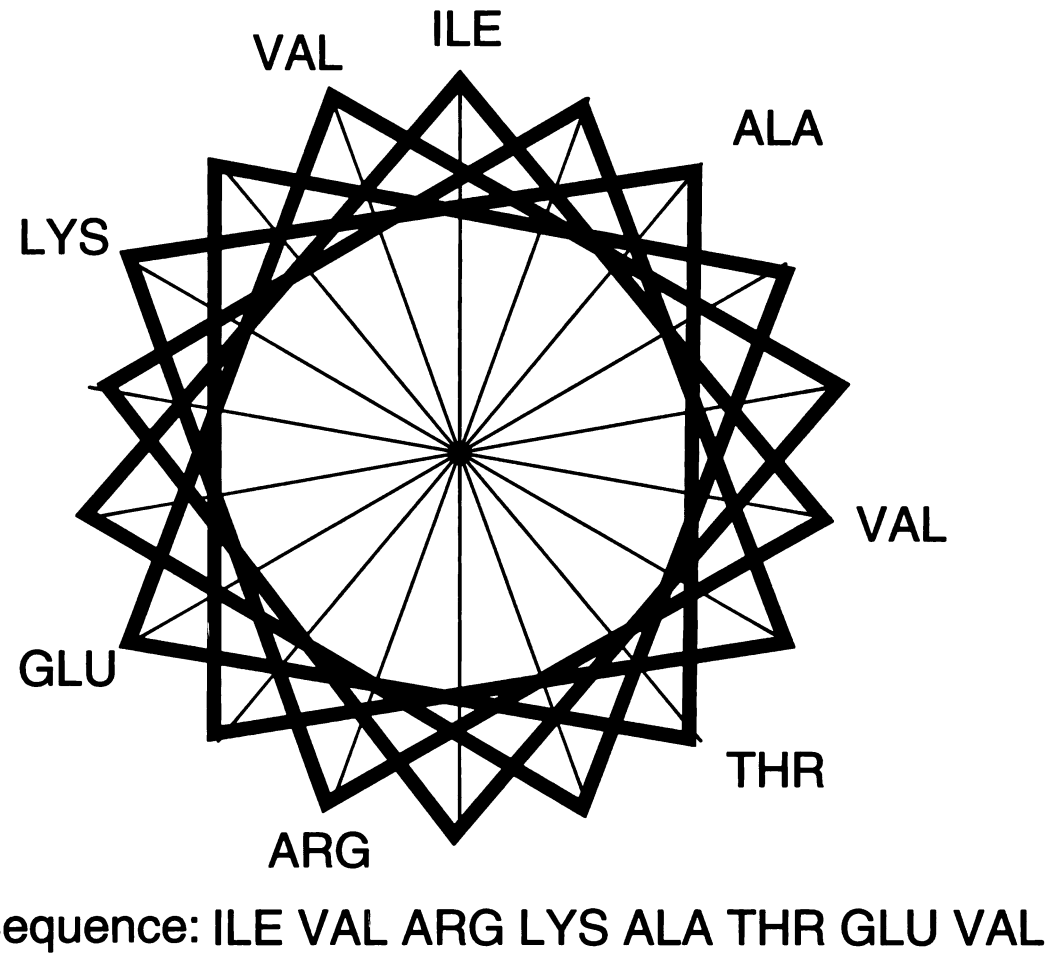


Figure II.4. Edmundson helical wheel. Formed by laying a sequence out in a circle, one amino acid every 100 degrees. In a helix, the hydrophobic and hydrophilic residues are grouped together. In this example VAL, ILE, and ALA are on one side, while LYS, GLU, ARG, and THR are on the other.

Appendix 1: Evaluation of Secondary Structure Prediction

Secondary structure prediction methods are rated by the percentage of amino acid residues correctly assigned by the method. The standard categories for secondary structure are α-helix, β-sheet, or other. A database of protein structures (the Brookhaven Protein Data Base is commonly used), is split into two parts: a training set and a test set. The training set is used to determine the parameters of the secondary structure determining algorithm. The test set is used to evaluated how well the algorithm works on a set on unrelated structures.

When secondary structure is assigned randomly to a set of sequences, and compared with the correct structures, 38% of the residues are predicted correctly. 33.3% accuracy would be expected for equal numbers of residues in each structure category, but the three states are not equally occupied.

The theoretical upper limit of secondary structure predictive accuracy is not 100%. Different crystallographers when asked to assign secondary structure agree approximately 85% of the time. Different algorithms used to analyze actual structures and assign their secondary structures agree about 85% of the time. The source of the discrepancy is differences in the precise definition of secondary structure.

A common variation on the three state model is the two state model (α-helix and other). The accuracy observed for a two state model will be higher than that of a three state model. Random choice of secondary structure would result in >50% predictive accuracy.

In many cases, the presence and location of general secondary structure units is more important than the exact point at which a helix ends or a long

UCSF LIBRARY

turn begins. Evaluation methods based on this idea have not been widely adopted, but are gaining in acceptance.

The final test for a method is to predict the secondary structure of a protein before the structure is published. Eventually, a database including the sequences of soon to be solved structures will be available to facilitate these predictions.

UCSF LIBRARY

predict secondary structure. These networks generate relationships amongst the different amino acids from an analysis of protein sequences and their structures. Neural Networks have achieved a 64% success rate. When trained on a set of exclusively helical proteins, networks correctly predict 80% of the conformational preferences of individual residues.

The limit in accuracy of secondary structure prediction is thought to be 65% for one sequence in the absence of long range (greater than 10 residues away) information. A set of homologous sequences that fold to approximately the same structure have been used to increase the accuracy of Neural networks to ~70%. Similar amounts of improvement have been shown with Garnier and Robson's and Chou and Fasman's methods.

Other biological information besides the sequence of the protein is also available. For specific proteins with several aligned sequences, Benner claims to be able to predict secondary structure with ~80% accuracy.

The structural class of the protein may also be used to increase secondary structure predictive accuracy. The secondary structure of proteins in the all helical class has been determined with an average of 80% accuracy. Increases in predictive ability have also been seen for the mixed α-helical and β-sheet class and for the all β-strand class. Part of the predictive improvement in the all α-helix class and the all β-sheet class is due to a reduction in the number of secondary structure types from three to two. For instance, the secondary structure predictive accuracy would improve for all proteins if the structure choices were just α-helix and other. The structural class of the protein may be determined by analysis of the amino acid composition of the protein. Zhang and Chou have determined whether or not a protein is all helical with ~100% accuracy for their test set. Experimental methods such as Circular Dichroism spectroscopy may help in the

USF LIBRARY

determination of structure class by allowing estimations of secondary structure in the protein.

More precise secondary structure elements are also predicted. Highly accurate (>90%) turn prediction algorithms can approximately assign the location of a turn. These algorithms may be used to improve the secondary structure prediction by finding the secondary structural element that is consistent with the distance between turns. Other programs have been designed to find the precise Nterminal and Cterminal ends of helices.

Tertiary Structure Prediction

The tertiary structure of a protein can be approximated by packing secondary structural elements together. These elements may be packedtogether in a myriad of different ways (**figure II.5**). Fortunately, the number of plausible tertiary structures is limited by constraints on secondary structure packing.

Two approaches are used to find the correct structure. The first is to manipulate the secondary structural elements on a graphics terminal. The biologist's knowledge of the protein, augmented by computational tools, is used to pack the protein into its probable structure. This structure is then evaluated by a number of tests of correct structure prediction. If errors in the structure are demonstrated, the structure is altered until it satisfies all the known constraints on its structure.

The second is to construct all the possible secondary structure packing arrangements (**figure II.6**). This group of structures is screened to remove structures that violate the known constraints on the structure. After the automated screens have been used, the remaining structures are evaluated by hand.

UCSF LIBRARY

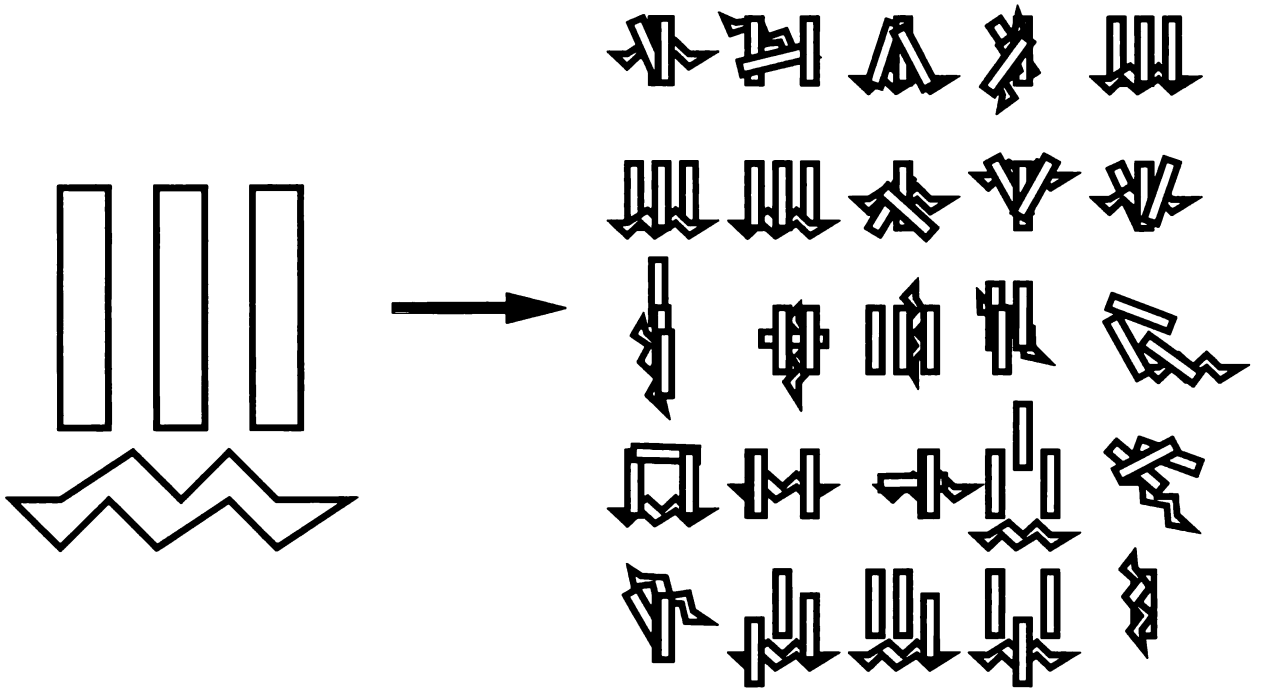


Figure II.5. Combinatorically arranged secondary structure elements. Three helices and a sheet can pack together in a myriad of different ways. This is a small subset of the possible conformations.

Figure II.6. Combinatorically arranged secondary structure elements screened according to compactness, solvation, and fluorescence measurements. In a real case ~100,000 structures, and as many as ten different screens would be used to generate and screen the possible structures.

Many methods have been found to evaluate protein structures. Each incorporates some methods into the design process while other methods are used to evaluate a generated structure. A list of constraints and methods may be found in **appendix 2.**

A structure that passes these tests can, in principle, be refined by detailed energy calculations. If a structure is determined by NMR or X-ray crystallography, the model structure may be evaluated. Methods for evaluation are given in **appendix 3.**

Worked Examples and Results: IL-4

The structure of interleukin-4 was calculated by the *de novo* approach and then compared to the subsequently determined NMR structure. Circular Dichroism spectroscopy was used to prove that IL-4 was dominated by α-helical structure. Secondary structure prediction methods were used to assign the location of α-helices and loops. A combinatorial algorithm generated all possible juxtapositions of the four helices subject to the constraints that a hydrophobic core was formed and that the interhelical loops could join neighboring helices. 90403 structures were generated that did not violate steric constraints or disrupt the connectivity of the chain. Of these, 311 were consistent with distance constraints imposed by the three disulfide bridges. Solvent accessible surface area calculations were used to select the energetically most sensible structures. When the three dimensional structure was solved by NMR spectroscopy, it was clear that the secondary structure was predicted accurately (~90%). Unfortunately, the best structure was the topological mirror image (**figure II.7**) of the NMR structure. The eighth structure on the list resembled the correct structure. (Root Mean Square Deviation = 4.8 Å)

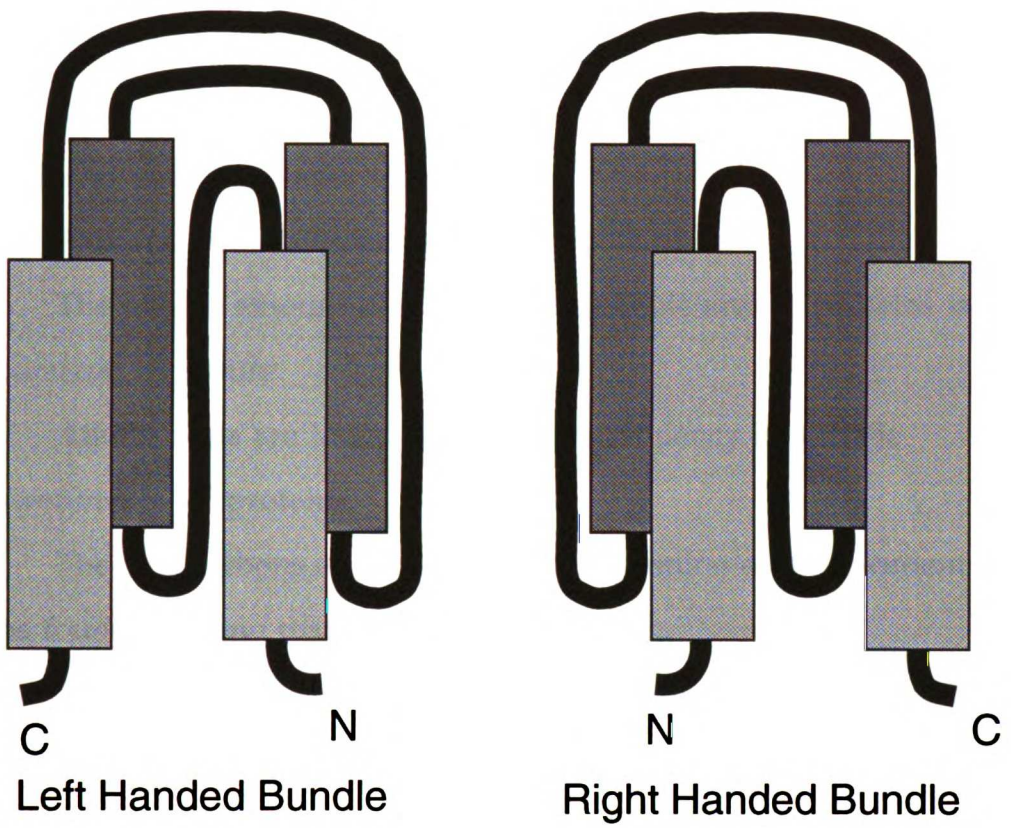


Figure II.7. The predicted structure of IL-4 and its correct topological enantiomer.

Appendix 2: Tertiary Structure Constraints and Screening Methods

Each secondary structural unit is constrained relative to at least one other unit by a connecting loop of amino acids.

The overall shape of the protein is limited by the known tendency of proteins to form a globular shape

Proteins form well-packed hydrophobic cores.

The overall amount of buried hydrophobic area in proteins is maximized in nature.

Amino acids are found with known frequency in specific environments in proteins.

The beta carbons of proteins exhibit an amino acid dependent tendency to be found a certain distance from one another.

The frequency and type of mutations in a group of related sequences is dependent on the environment of each amino acid.

Simultaneous mutations in a group of related sequences may be indicative of amino acids close in physical distance.

A disulfide bond severely restricts the distance between two cysteine residues.

The properties of mutant proteins may suggest amino acids key to protein stability or resolve a specific structural conflict.

Tryptophan fluorescence may be used to determine the accessibility to solvent, and freedom of movement, of tryptophan residues.

Alternate fluorescent probes may be included in the protein sequence to examine the environment of a specific residue.

Charged amino acids have relative spatial distributions dependent on their charge and environment.

LIBRARY

Appendix 3: Tertiary Structure Evaluation

The accuracy of a tertiary model is measured against the actual protein structure. Often the distances between corresponding atoms in the structures are compared. The lowest possible root mean square distance between the structures is called the RMS difference. Usually just the a-carbons are used, but sometimes all the atoms of the protein are used.

The RMS score is also used to describe the differences between the intra-atomic distances of one structure with the intra-atomic distances of another. These scores are usually lower than the standard RMS and can lose some details of the model.

Both RMS measures are sequence length dependent. Random sequences compared to a 60 a.a. structure have an average RMS (standard method) of 6.9 Angstroms. The RMS for a 250 a.a. structure is 12.43 Angstroms.

A structure sometimes has a relatively low RMS score while having fairly incorrect overall topology, while another has the correct topology but a somewhat high RMS score. In addition, the non-standard RMS model cannot be used to differentiate between structures that are mirror images of one another.

Protein models are often visually compared to the correct structure due to a lack of a satisfactory method for evaluating structures.

UCSF LIBRARY

Modeling By Homology

Sequence Alignment

Modeling by homology follows a four step recipe: sequence alignment, framework construction, loop construction and side chain placement. These steps are detailed in **figure II.8**. In the first step the sequence of the protein is aligned with the sequence(s) of a protein(s) of known structure. A sequence alignment with greater than 50% residue identity is desirable but 20-30% can be used if several aligned sequences and structures are available. The sequence alignment is crucial to the development of an accurate model. Programs have been developed to align a sequence to a structure that are superior to alignments of sequences to other sequences. The alignments still need to be modified in the framework stage to reflect precise structural information the alignments have not captured. When multiple structures are available, inconsistencies in the alignment can be resolved by comparing to both structures. This allows the lower resolution structures to be used.

Framework Construction

After the alignment is made, a portion of the new protein structure is constructed from the core secondary structure of the known structure. If the structures of several homologous sequences are available, a representative framework or an aggregate framework structure can be used. The core is built first since it is the least likely to deviate from the template structure.

Loop Modeling

Helices and sheets are comparatively easy to model for two reasons: they conserve their local conformation and their location relative to other secondary structures across evolutionarily related proteins, and they contain repeating geometries. This is not true for the aperiodic loop regions that join

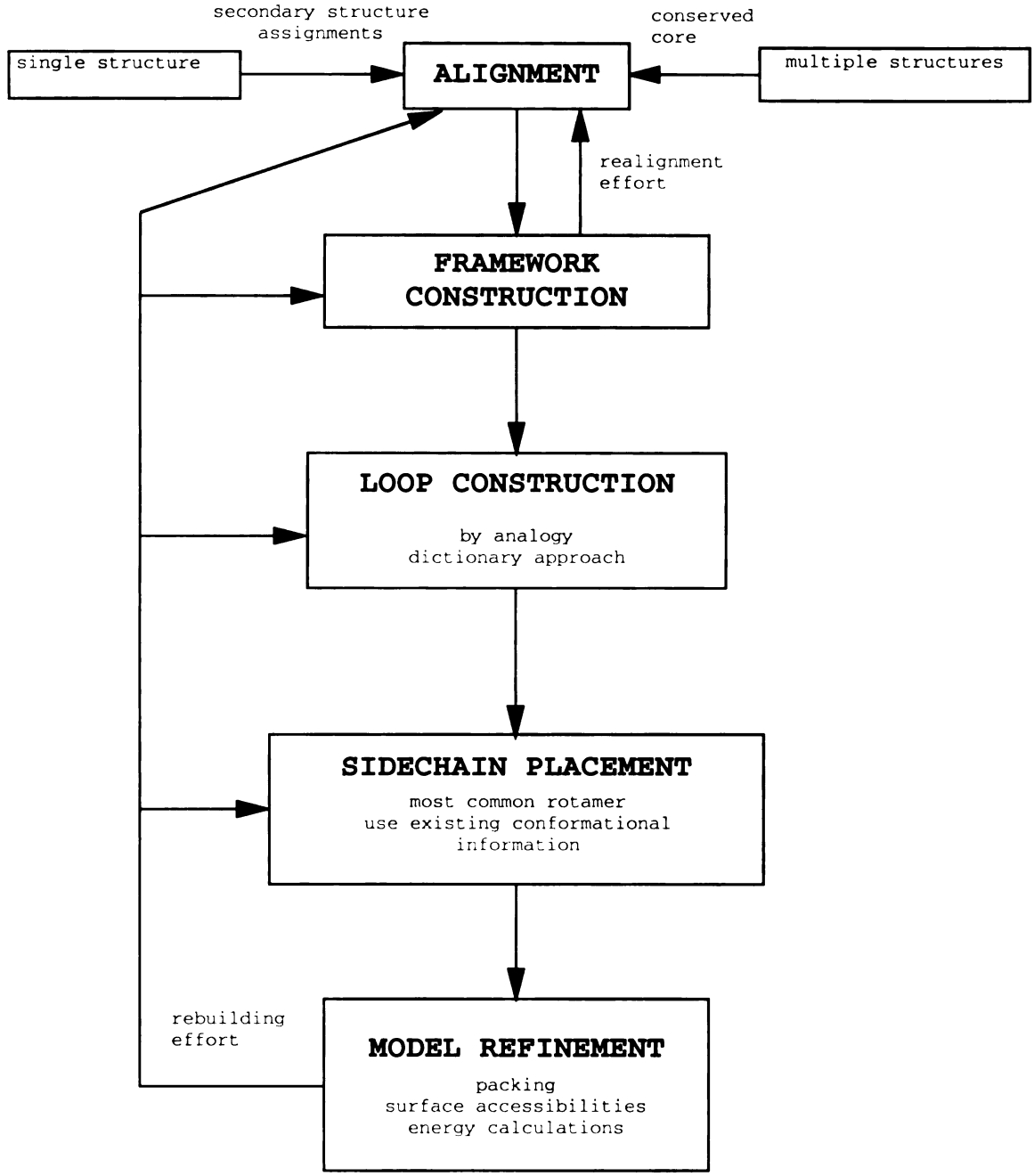


Figure II.8. Flow chart for modeling by homology. If an error in the structure is detected, the structure is rebuilt from an earlier phase to correct the error. Most errors are found in the framework construction or model refinement phase.

secondary structure units. Thus, several methods have been devised to model loop structure. The simplest and most effective method for anticipating the conformation of a loop extends the notion of framework homology to loops with sequences that are very similar to a corresponding loop of known structure from the family of homologous structures. If the loop is a member of a well-characterized set such as the β-turns, a four residue transition between two anti-parallel β-strands, it is possible to choose the β-turn consistent with the sequence and distance constraints. Otherwise, one exploits a dictionary of loops derived from all proteins of known structure assembled in the PDB. The loop fragment from the PDB provides a plausible conformation for the region of structure under study. This is most effective with short loops (length 2-10 amino acids) since longer loops are under represented in current versions of the protein database.

The *ab initio* method is used to select the lowest energy loop conformation from all the possible loop conformations. This approach is computationally difficult and intractable for loops that contain more than seven amino acids owing to the vastness of conformational space.

Chothia and coworkers have characterized immunoglobulin loops and found that they are confined to a limited set of canonical structures. In this case, the correct loop conformation may be chosen from this set. This was possible do to the large number of immunoglobulin structures available.

Side Chain Placement

Most protein side chains are placed in a conformation that is reminiscent of the side chain geometry adopted by the homologous residue. Summers et al. found that 80% of identical residues and 75% of mutated residues retained their side chain conformations when the proteins shared greater than 40% similarity.

When the side chain geometry is not known, the residue is placed in its statistically most likely conformation. This is the case when the mutated residue was a glycine or proline, resulted from n amino acid insertion, or is in a mutated loop. The specific rotamers of the side chains and their statistical likelihood have been compiled from the PDB. These rotamers are found to occupy specific discrete angles. Recently, the rotamer probabilities have taken into account neighboring amino acids. This information may be used to assign rotamers to side chains of unknown conformation. It is also possible, though computationally difficult, to search through many side chain rotamers simultaneously and evaluate them according to a packing and/or energetic consideration.

Model refinement and Structure Validation

Energy calculations, including energy minimization and molecular dynamics, are used to locate a structure with sensible steric and electrostatic interactions. Unfortunately, energy calculations are not sufficiently accurate to confirm that the model built structure is correct. As a result, models are analyzed by a series of empirical measures. These measures are similar of the same as those used to evaluate *de novo* models (**appendix 2**). Efficient packing, the integration of the hydrophobic core and appropriate residue solvent accessibility profiles are some of the methods used. If part of a model appears to be inconsistent with these empirical measures, a rebuilding effort is required (**figure II.4**).

Worked example and results: Malarial protease

Homology model built structures have been shown to be accurate enough to aid in a drug discovery program. These structures are accurate enough to identify the basic shape of the active site and to search a database of

small molecules to find the closest fit. This approach was used to identify a possible drug for the treatment of malaria. Malaria feeds by breaking down hemoglobin with an enzyme called a protease. Two protein structures were available that had a high degree of homology to the malarial protease. A model for the malarial protease was constructed using this homology modeling method. The active site was known from previous studies, so a database was searched to find a structure that fit this active site using a program called Dock developed by Tack Kuntz. The database was also searched to find structures that were electrostatically compatible with the active site. These two lists were combined and visually edited to produce a subset of testable compounds. A few of these inhibited the protease in a test tube, and one was effective against the malarial parasite. An ongoing effort is being made to refine the drug to be effective in a living system.

The model structure may have succeeded for two reasons. First, the malarial protease was successfully modeled. Second, the template structure was sufficient to find the inhibitor and the modeling effort was unnecessary. The inhibitor of the malarial protease was not effective against the template protein. When Dock was used on the template protein, the best inhibitors of the malarial protease did not appear on the list of possible structures. These two results proved the first hypothesis, for this case. However, in another case, a homology built model was found to be farther from the subsequently determined crystal structure than from the template protein.

New approaches

The initial step in homology modeling is to find a sequence that has homology to the sequence of a structure, or find a structure that has sequence homology to a sequence. This is normally done by direct sequence comparison. However, sequences are known that are quite dissimilar from

one another, yet fold to the same overall structure. Methods have been developed to find homologous structures when the sequence homology between them is quite low.

An early approach involved finding a consensus sequence that represented a group of protein structures. This consensus sequence was more successful than a single sequence for identifying structurally related proteins. A subsequent approach simplified the protein structure to a one dimensional string of environments. The environments of amino acids in a subset of the PDB were tabulated, and converted to preferences of amino acids for environments. This allowed the compatibility of a sequence with a string of environments to be measured. Recent methods string sequences on known protein structures. The distance between the b-carbons is tabulated, and the model is evaluated based on the known distance separation of amino acids in the PDB. These techniques have allowed structurally related proteins to be found that have minimal sequence similarity.

As more and more protein structures are solved by X-ray Crystallography and NMR, the importance of the structure homology methods increase. Chothia has estimated that there are only 1000 distinct protein families. ~150 of these families has a representative in the PDB. Eventually, each new sequence determined will have a high chance of resembling a previously solved structure, lessening the need for future *de novo* and crystal structures.

Molecular Dynamics

Energy Functions and Molecular Mechanics

The attraction or repulsion each atom feels for every other atom can be described as a sum of interaction energies. The functional form of this expression is derived from structural studies of proteins and small molecules as well as from theoretical and thermodynamic studies. A typical molecular mechanics potential takes the form:

$$E(\vec{x}_1,\ldots\vec{x}_m) = \varepsilon_{bond} + \varepsilon_{ang} + \varepsilon_{tor} + \varepsilon_{vdw} + \varepsilon_{el}$$

$$E = \sum_{i=1}^{m} K_b(r_i - r_b)^2 + \sum_{i=1}^{m} K_a(\theta_i - \theta_a)^2 + \sum_{dihedrals} \frac{K_d}{2}[1+\cos(n\phi - \chi)] + \sum_i \sum_{j>i} (B_{ij} r_{ij}^{-12} - A_{ij} r_{ij}^{-6}) + \sum_i \sum_{j>i} \frac{q_i q_j}{\varepsilon r_{ij}}$$

E is the total energy of the system.

ε_{bond}, ε_{ang}, ε_{tor}, ε_{vdw}, ε_{el} are components of the total energy representing bond, angular, torsional, van der Waals and electrostatic energy respectively.

K_b, K_a, K_d are force constants associated with bond, angular and torsional energies, respectively.

r, r_b are the bond distance and the equilibrium bond distance, respectively.

q, q_a are the bond angle and the equilibrium bond angle, respectively

n is the periodicity of rotation.

ϕ is the dihedral angle.

χ is the phase angle of the dihedral angle.

i, j represent different atoms in the protein.

r_{ij} is the radius between atoms i, j.

A_{ij}, B_{ij} are the nonbonded (Lennard-Jones) repulsion and attraction coefficients for the interacting atoms i, j.

q_i, q_j are point charges of the atoms i, j.

ε is the dielectric associated with the molecular environment.

This function is useful for revealing incorrect contacts or charge arrangement in proteins. This potential does not accurately describe the energies of components of the system that have a high degree of mobility. This includes the side chains, and the sovlent surrounding the protein. These components can be modeled in a dynamic system.

Equations of Motion

Molecular dynamics is the study of how the molecular potential changes with time given an initial set of atomic positions and velocities (temperature). The motion of the atom in a protein can be described by Newton's equation of motion:

$$F = -\nabla E(\vec{x}_1, \ldots, \vec{x}_m)$$

$$F = \frac{m\partial^2 x(t)}{\partial t^2}$$

F is the force on an atom.

E is the energy of the system.

$x_m \cdots x_m$ are vectors representing the psational coordinates of all the atoms in the system.

m is the apporpriate atomic mass.

t represents time.

$x_i(t)$ represents the position of one atom in the system at a given time.

With an integration time step of 1-2 femtoseconds, these equation are well behaved.

What Can You Learn?

Computational constraints limit dynamic simulations to the nanosecond time frame. This is too short for many biologically important conformational changes such as protein folding which generally occurs in microseconds. Still, many useful applications have been devised.

Molecular dynamics has been used to follow the movement of molecular oxygen toward the heme iron of myoglobin. X-ray crystal structures are refined with molecular dynamics . Crystallographers resolve a protein structure form the diffraction pattern exhibited by X-rays that have passed through a protein crystal. At one stage of this process, a hypothetical diffraction pattern is generated from a derived model of the protein. This diffraction pattern is compared with the true diffraction pattern. The atoms in the model are then caused to move with molecular dynamics in an effort to find a structure that more closely satisfies the observed diffraction pattern. This approach can greatly improve the accuracy of crystallographically determined structures.

Molecular dynamics can also be used to calculate absolute free energy values for simple systems. Most modeling techniques only allow the

calculation of a pseudo energy value with no physical counterpart. Absolute free energies do represent a physical property and can be used to calculate rates of reactions and equilibrium constants.

The crystal structure of a lead compound complexed with an enzyme can be examined in this way. Free energy calculations are used to analyze the effect various small chemical alterations would have on the drug's effectiveness, simplifying the drug optimization process. Drug design is being attempted by finding the optimal location and composition of small chemical groups on a protein surface. These groups can be assembled to form a lead compound for synthesis. Alternatively, a databases may be searched for compounds with a similar arrangement of groups.

Simplified Systems

Simplified energy function are used to study properties of proteins that are apparent on longer time scales. These simplified models are designed to remove some of the complexities of the system while still mimicking specific properties of the protein, such as secondary structure content. Dill et al. have examined the behavior of simplified proteins confined to two or three dimensional lattices. The atoms of the protein are allowed to be discrete distances apart and in discrete orientations. For small peptides, all the possible orientations of the protein chain can be enumerated. These simulations have given insight into the important stabilizing effects in proteins, and into the properties that lead to secondary structure formation. Analysis of this work suggests that the compactness of proteins is sufficient to form the secondary structure observed in real structures. This is contrary to the popular view that hydrogen bonding is almost solely responsible for the observed secondary structure. An off-lattice simple model of proteins constructed and analyzed by Cohen et al. indicates that compactness does

indeed lead to secondary structure, but not in the amounts found by Dill. Other models have been used to examine the denaturing effects on proteins of high temperature, low temperature and various solvents. It is hoped that future models will offer still more insight into the properties of protein folding.

References

Allen, M. P. & Tildesley, D. J. (1989). *Computer Simulations of Liquids*, Clarendon Press, Oxford.

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

Brandon, C. & Tooze, J. (1991). *Introduction to Protein Structure*, Garland Publishing, New York.

Brooks, C. L., Karplus, M. & Pettitt, M. (1988). *Proteins - A Theoretical Perspective of Dynamics, Structure and Thermodynamics*, John Wiley & Sons, New York.

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

Fasman, G. D. (1989a). *Prediction of Protein Structure and the Principles of Protein Conformation*, Plenum Press, New York.

Fasman, G. D. (1989b). Protein conformational prediction. *Trends in Biol. Science.* **14**, 295-9.

Garnier, J. (1990). Protein Structure Prediction. *Biochimie* **72**, 513-524.

UCSF LIBRARY

Kuntz, I. D. (1992). Structure-based strategies for drug design and discovery. *Science* **257**(5073), 1078-82.

Lesk, A. M. (1991). *Protein Architecture: A Practical Approach*, IRL Press, Oxford.

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

Chapter 3

Structure of the A domains of factor VIII determined by homology modeling

UCSF LIBRARY

This chapter has been published with co-authors Yang Pan, Fred Cohen and Jane Gitschier in Nature Structural Biology, Volume 2, pages 740-44, 1995

We have predicted a structure for the three A domains of the blood coagulation factor VIII, which comprise the bulk of activated factor VIII, by virtue of their homology to blue-copper binding proteins of known structure. Each A domain is composed of two β-barrels, and the three A domains are arranged in a triangular configuration. The model agrees with our prediction of a type-II copper binding site linking the A1 and A3 domains. The debilitating effects of 84% of reported missense mutations associated with severe hemophilia A can be explained by the model.

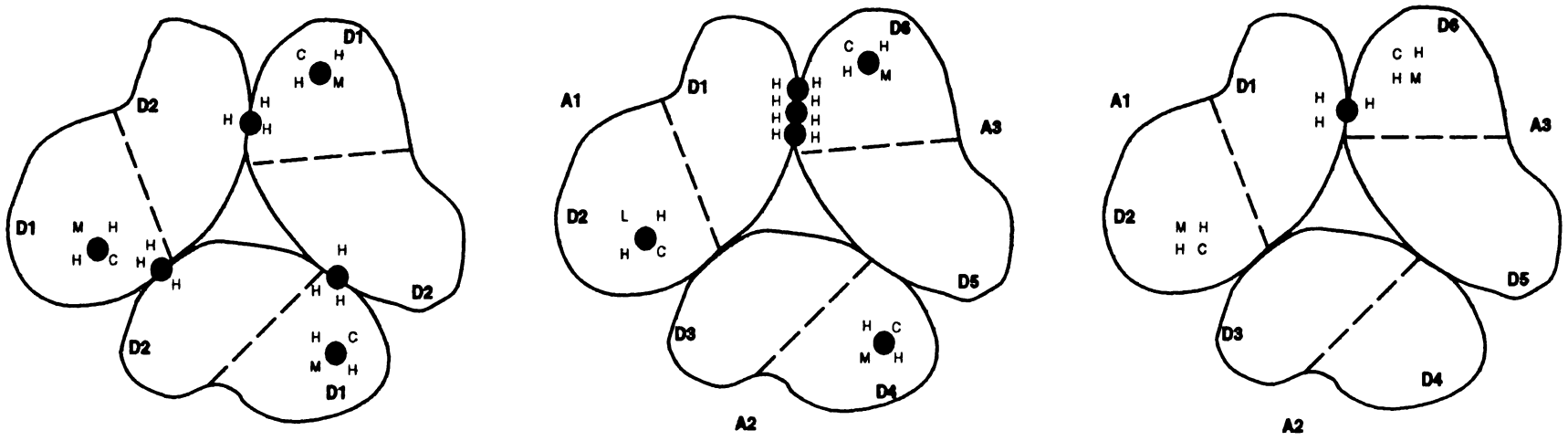
The X-linked bleeding disorder hemophilia A, affecting 1 in approximately 5000 males worldwide, is caused by mutations in the gene encoding coagulation factor VIII. As a cofactor in the intrinsic coagulation pathway, factor VIII increases the Vmax of activated factor IX in the conversion of factor X to Xa. Study of the mechanism by which factor VIII serves its essential cofactor function has been limited by a lack of structural information. Its large size (2351 amino acids) and low abundance (< 1nM in the plasma) have hindered the study of its structure by crystallography or NMR spectroscopy.

The primary structure of activated human factor VIII consists largely of three A domains (~330 amino acids each) on separate polypeptides, generated by cleavage from a much larger single chain precursor (Vehar *et al.*, 1984). Two small C domains (~150 amino acids each) are joined to the carboxy terminus of the A3 domain and three short acid amino acid-rich regions (~40 residues) are located at the carboxy termini of the A1 and A2 domains and the amino terminus of the A3 domain. Dissociation or cleavage of the A2 domain results in inactivation of factor VIII (Eaton *et al.*, 1986; Fay & Smudzin, 1992; Pittman *et al.*, 1992). A domains are also found in a related blood coagulation cofactor, factor V (Kane & Davie, 1988), and are approximately 40% identical to the triplicated A domains constituting the circulating copper-binding protein ceruloplasmin (Takahashi *et al.*, 1984).

A β-barrel motif has been observed in a large number of phylogenetically distant blue copper-binding proteins such as plastocyanin (Collyer *et al.*, 1990; Guss & Freeman, 1983), pseudoazurin (Petratos *et al.*, 1987), azurin (Baker, 1988), ascorbate oxidase (Messerschmidt *et al.*, 1992)(AOZ) and nitrite reductase (Godden *et al.*, 1991)(NIR). NIR and AOZ monomers consist of two or three β-barrel containing domains, respectively, while others have only a single β-barrel. Previously detailed comparison of ceruloplasmin with AOZ and plastocyanin indicated that each of its A domains can further be divided into two subdomains and that each subdomain probably forms a single β–barrel unit similar to those found in blue copper binding proteins (Dwulet & Putnam, 1981; Messerschmidt & Huber, 1990; Ryden, 1982). By analogy, the factor VIII A domains should also divisible into two subdomains, suggesting there are six β-barrel containing domains (designated as D1 to D6) in the activated factor VIII.

For modeling purposes, the protein with known structure which most closely approximates the proposed six β-barrel structure of the A domains of activated factor VIII is NIR, a homotrimer with two β-barrels in each monomer (Godden *et al.*, 1991). We have used the trimeric structure of NIR as a framework for modeling the pseudo-trimeric A domains of factor VIII, extending a previous proposal for the gross structural arrangement for ceruloplasmin A domains (Fenderson *et al.*, 1991), as described in **figure III.1**.

Figure III.1 also shows the amino acid alignment of the putative β-barrel domains (D1 to D6) of factor VIII and ceruloplasmin with those from the two most closely related blue multicopper-binding proteins, AOZ and NIR. For factor VIII and ceruloplasmin, the amino terminal β-barrels of each A domain (D1, D3, D5) group together because they are more related to each other than they are to the carboxyl terminal β-barrels (D2, D4, D6) and vice versa. Structural information from AOZ, NIR, plastocyanin, pseudoazurin and azurin was used to derive a sequence



Nitrite Reductase

Godden et al., 1991

Ceruloplasmin

proposed by

Fenderson et al., 1991

Factor VIII

(a)

Figure III.1. Comparison of factor VIII A domains to nitrite reductase and cerruloplasmin. *a*, Proposed structural arrangement which preserves the type I copper-binding sites within corresponding D2, D4 and D6 domains, and the type II copper-binding site at the domain interface. *b, next page,* Amino-acid sequence alignment of the six domains from factor VIII and ceruloplasmin, the three domains from AOZ and the two domains from NIR. Secondary structure of β-strands from known AOZ and NIR structures are underlined and alphabetized. Ligands forming putative type I, II and III copper-binding sites are coloured green, blue and pink, respectively. Cysteines involved in putative disulphide bridges are indicated in yellow. Factor VIII and ceruloplasmin sequences were taken from GenBank and coordinates for AOZ from zuchini (*Cucurbita pepo medullosa*), NIR from *Achromobacter cycloclastes,* azurin from *Alcaligenes denitrificans,* plastocyanin from *Enteromorpha prolifera,* and pseudoazurin from *Alcaligenes faecalis* were obtained from the Brookhaven Protein Data Bank (Bernstein *et al.*, 1977).

(b)

alignment for factor VIII. Knowledge of type I copper binding ligands and the precise location of β-strands in NIR, AOZ and other related proteins was used to focus alignment efforts on the structurally conserved regions (Greer, 1990). When two or more plausible alignments emerged, tertiary structure preferences were considered leading to refined alignments in an iterative fashion.

One important structural feature in these blue copper proteins involves the amino acid ligands for copper-binding associated with distinguishable spectroscopic properties. In type I, the "blue" copper is liganded to four residues: sequentially His, Cys, His and Met. The ligands for types II and III coppers are formed by two or three histidines. Azurin, pseudoazurin and plastocyanin contain only one type I copper in their one β-barrel structure. NIR, AOZ and ceruloplasmin contain multiple copper-binding sites. Each NIR monomer has one copper liganded in a type I site, and an additional type II copper is bound by three histidines at the interface of each of two monomers within the homotrimer. Specifically, the type II copper-binding sites are formed by His100 and His135 of the D1 domain of one monomer together with His306 of the D2 domain of the next monomer (these three histidines are labeled blue in **Figure III.1**). The sequence comparison with AOZ suggests the existence of three type I, one type II, and two type III copper binding sites (Messerschmidt & Huber, 1990) in ceruloplasmin (**Figure III.1**), similar to those found in AOZ. This prediction agrees with the experimental data on copper type and content in ceruloplasmin (Frieden, 1980). Our alignment indicates that factor VIII also has two potential type I copper-binding sites, although there is no experimental evidence to suggest that these sites are occupied by copper.

Inspection of the amino acid sequence alignment reveals a possible type II copper-binding site in factor VIII. Two of the three histidines involved, His99 and His161, are from the A1 domain (D1) and the third, His1957, is from the A3 domain (D6). (These three histidines are labeled blue in **figure III.1**.) At the equivalent

positions within the other domains of factor VIII, the triad of histidines is lacking. We hypothesize that the putative type II copper-binding site is used in factor VIII to tether the A1 and A3 domain together, and that lack of such a site at the A1/A2 interface and A2/A3 interface could explain why the A2 domain is more readily dissociated from the rest of the complex.

The D1, D3 and D5 domains of factor VIII were modeled after the D2 domain of the NIR, because the type I copper-binding site is common to all of them and these domains exhibit the greatest degree of overall similarity. Conversely, most of D2, D4 and D6 domains were modeled after the D1 domain of the NIR due to greater sequence similarity. As a result, the first and last residue of each A domain are in adjacent positions. The coordinates for the C-terminal portion of these domains were adopted from the corresponding portion of AOZ, due to greater similarity. Between any D domain sequence of factor VIII and related sequences with known structures, the percentage of amino acid identity ranges from 11% to 28% and similarity ranges from 38% to 51%.

The factor VIII and ceruloplasmin sequences were taken from GenBank and Coordinates for ascorbate oxidase from zucchini, nitrite reductase from *Achromobacter cycloclastes*, azurin from *Alcaligenes denitrificans*, plastocyanin from *Enteromorpha prolifera* and pseudoazurin from *Alcaligenes faecalis* were obtained from Brookhaven Protein Data Base (Bernstein *et al.*, 1977). Insight II, a package of protein homology modeling program from Biosym was used for the modeling and evaluation (Insight II User Guide, version 2.3. San Diego: Biosym Technologies, 1994). Following the sequence alignment, the sidechains of the NIR D domains were replaced with those of factor VIII in the structurally conserved regions using the approach of Summers and Karplus (Summers *et al.*, 1987). Steric clashes generated by sidechain replacement were relieved by rotating the sidechains of the close packed residues consistent with the desire to fit side chain dihedral angle

preferences (Dunbrack & Karplus, 1993). All loops were computer generated because no comparable loops from known structures could be found. The packing density was examined with QPACK (Gregoret & Cohen, 1990) which identified severely over-packed and under-packed areas. Solutions to these structural problems were sought by varying the relevant side chain rotameric states. The final model was energy minimized with the cvff forcefield in Insight II. Careful attention was paid to the structure to avoid over-minimization and subsequent structural distortions. Ribbon drawing and space filling model were generated in program MIDASPLUS (UCSF graphics facility).

The resulting structure of the factor VIII A domains is shown in **Figure III.2a**. The model consists of six β-barrels, roughly contained in a sphere with a radius of 80 Å. It includes 987 amino acids for the three A domains of factor VIII. Although the acidic domains are not a part of this structure, they are projected to be on the same side of factor VIII (the side facing the viewer).

Four lines of evidence suggest that this model is approximately correct. First, six disulfide bonds are predicted in factor VIII by this model, even though equivalent disulfide bonds do not exist in the structure of reference protein NIR. Ten of twelve cysteines predicted to participate in disulfide bond formation (C153-C179; C248-C329; C528-C554; C630-C711 and C1832-C1858) are conserved in both factor V and ceruloplasmin. Although these five disulfide bridges were not confirmed directly in factor VIII, their counterparts in factor V have been determined (Xue *et al.*, 1994). **Figure III.2b** shows an enlargement of the structure of the region containing two of the disulfide bonds in the A2 domain. The distances between two alpha carbons of each pair are ~ 5 and 7 Å respectively. An additional disulfide bond was also predicted between C1899 and C1903 between β-strands a and b in the D6 domain.

Second, the model reveals that the three histidines (His99, His161 and His1957), which we predicted could be involved in formation of type II copper

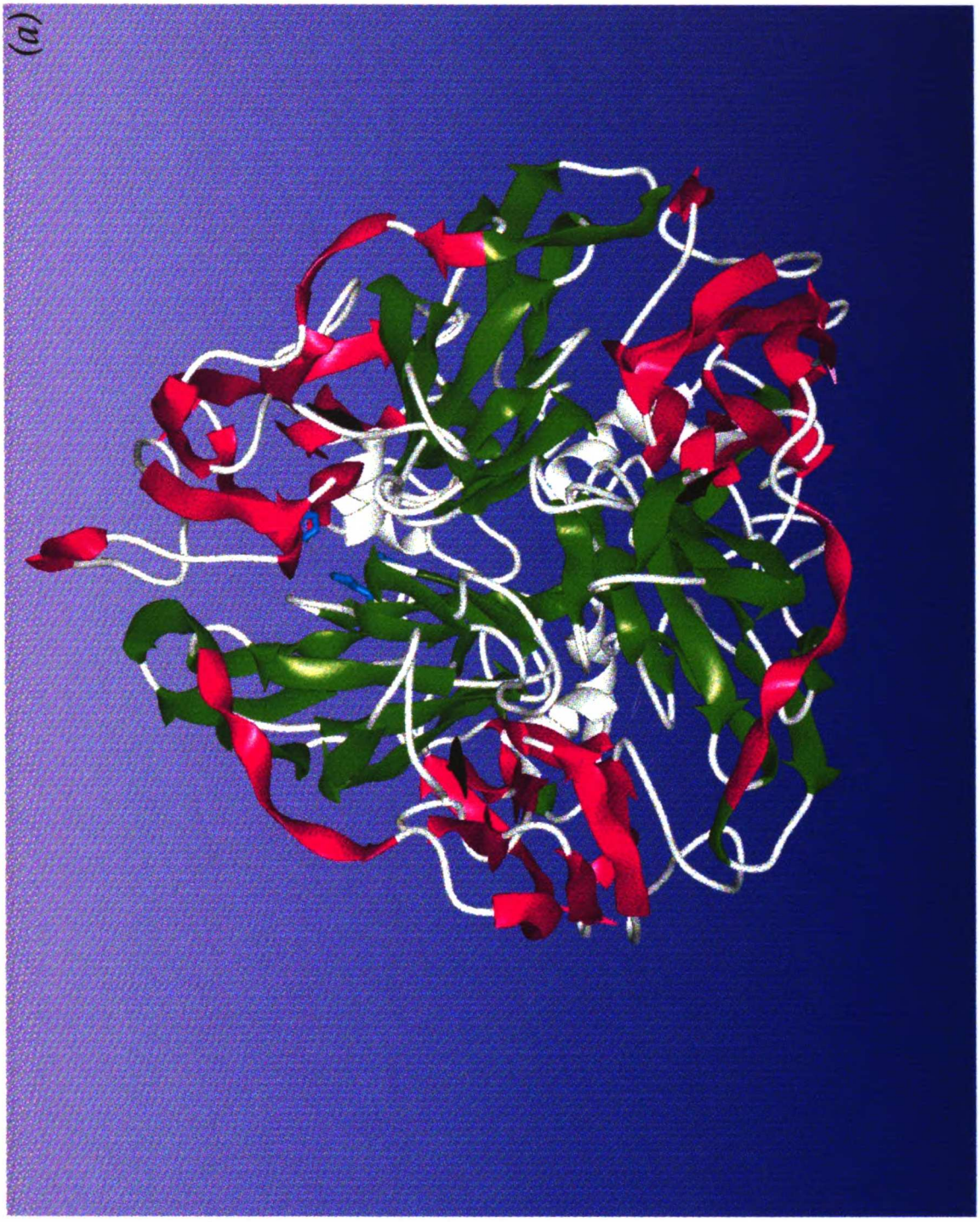
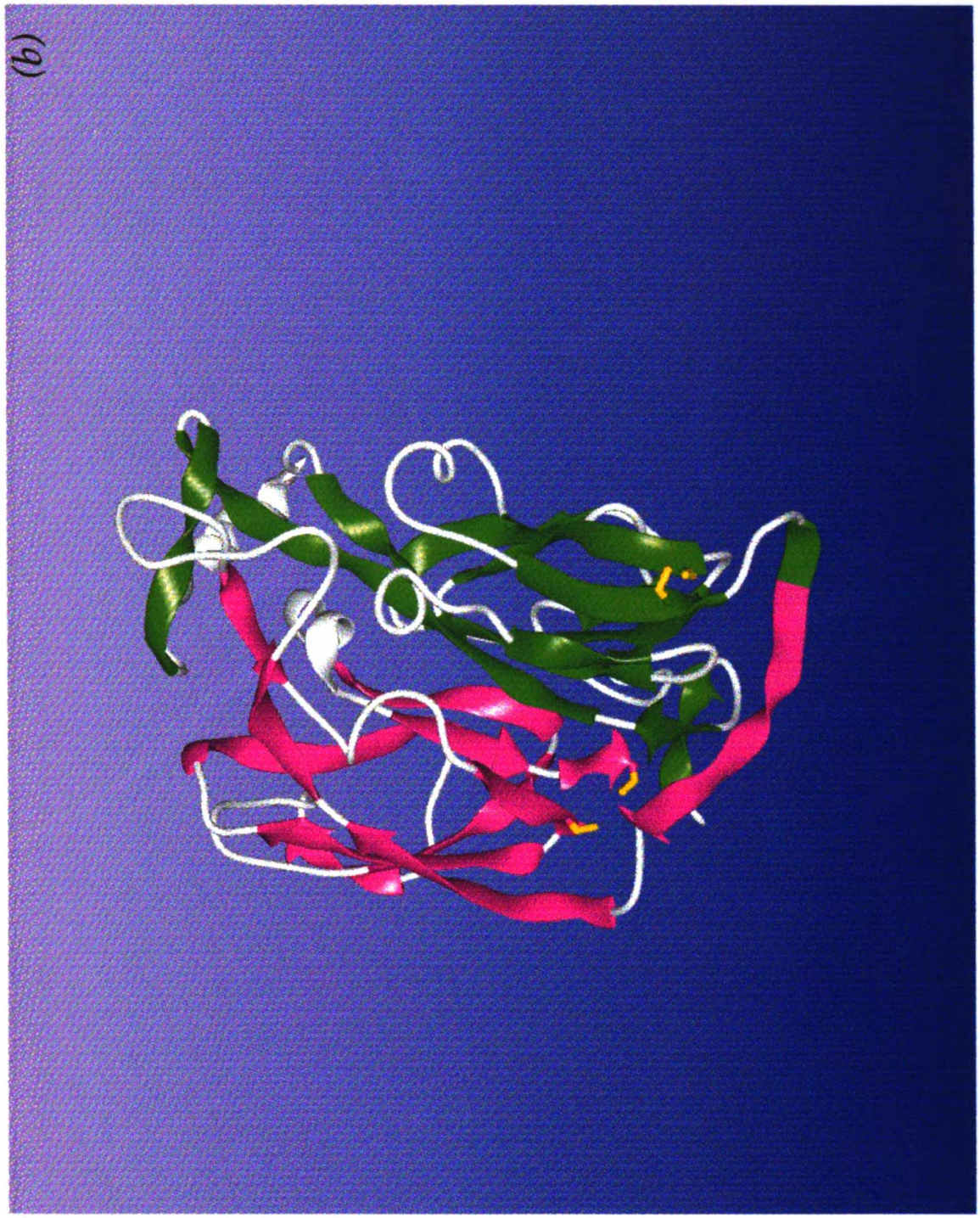
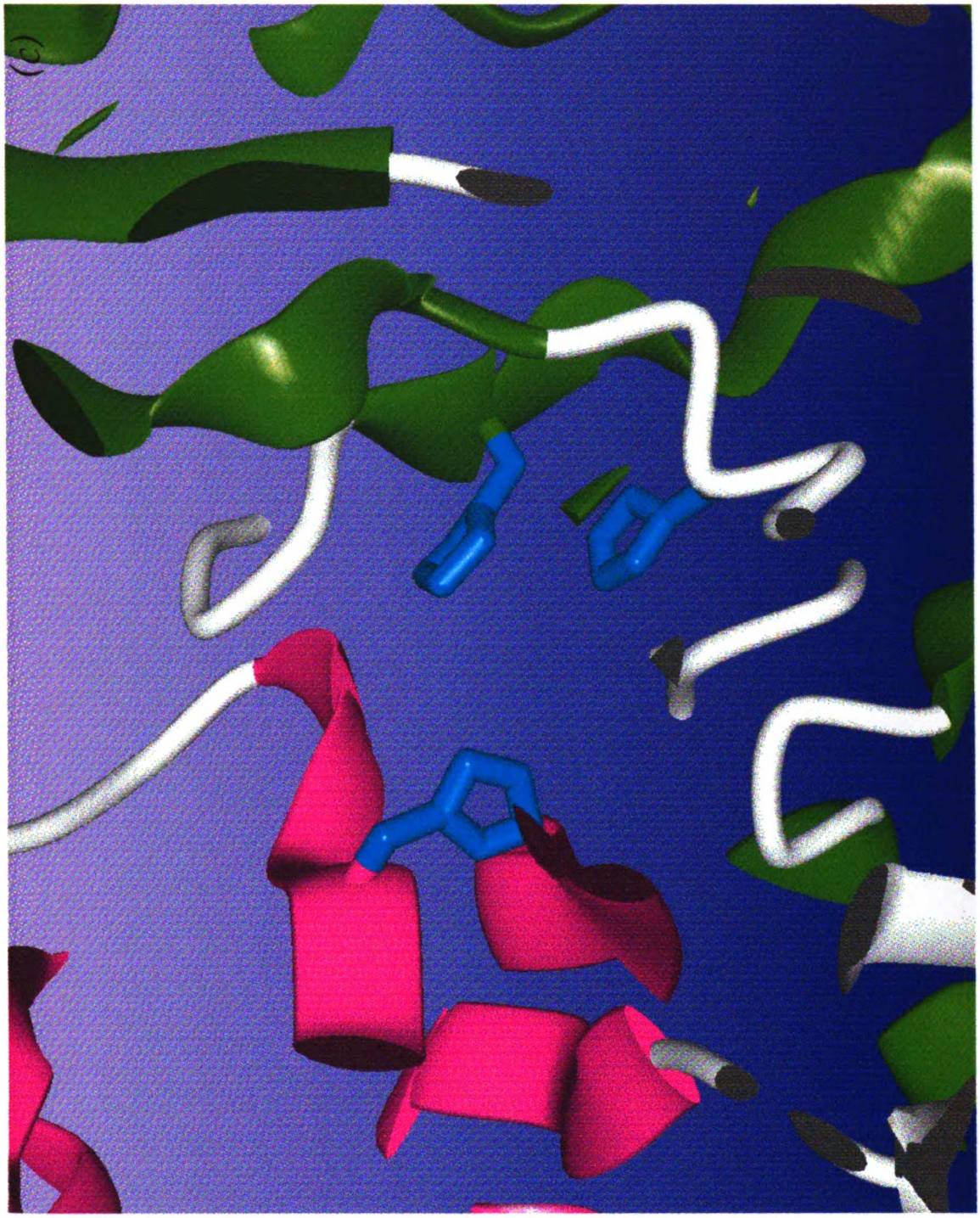


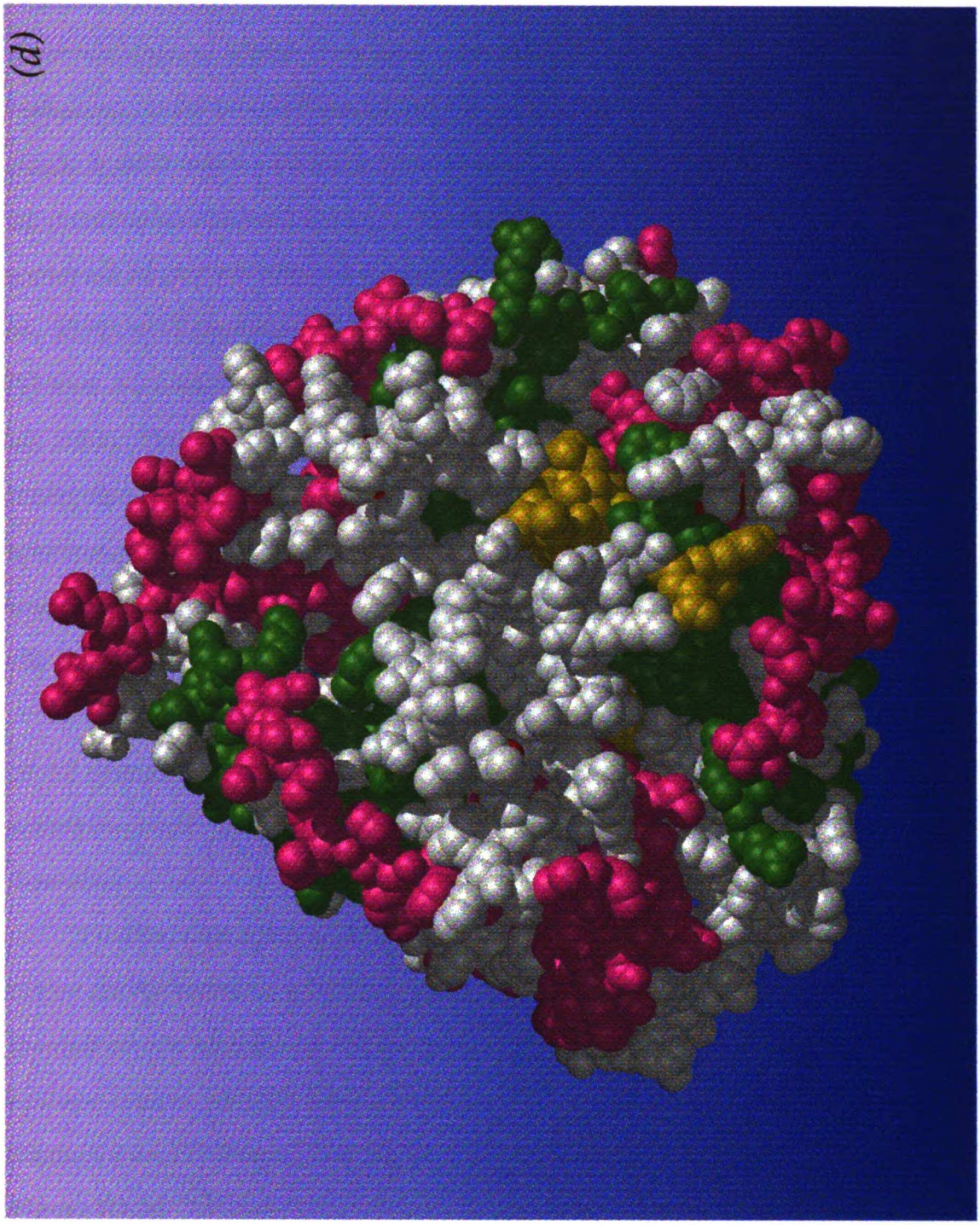
Figure III.2. Structure of factor VIII A domains by homology modeling. *a*, Ribbon structure of the three A domains, with arrows indicates β-strand and spirals indicate helices. The b-strands in D1, D3 and D5 domains are colored green. The β-strands in D2, D4 and D6 are colored magenta. The rest of molecule is colored white. Three histidines involved in copper-binding are highlighted in blue. The last residue for each A domains is colored black.
(*continued next page*)



b, Structure of A2 domain with disulfide forming cysteines shown in yellow. This structure is shown in an orientation different from that in a for better viewing of the cysteines. The activated protein C (APC) site is located at the junction of the two β-barrels.
(*continued next page*)



c, Enlargement of the type II copper binding site, viewed from the back of the structure shown in *a*, with three histidines colored blue.
(*continued next page*)



d, Spacefilling model of the three A domains, shown in the same orientation as in *a*, indicating two regions (shown in gold) which are implicated in inhibitor binding, residue 389 to 391 and 485 to 509.

binding, are situated in close proximity at the interface of the A1 and A3 domains, as shown in **Figure III.2c**. Together with a water molecule (as in NIR), these three histidines could be the ligands for a type II copper binding site. This prediction of a single type II copper-binding site, connecting the A1 and A3 domains, is supported by recent experimental studies. One copper ion has been found to be present in one molecule of plasma-derived factor VIII by atomic absorption spectroscopy (Bihoreau *et al.*, 1994), and thedissociation of this copper ion coincides with the dissociation of the A1 and A3 domains in both plasma-derived and recombinant factor VIII (Bihoreau *et al.*, 1994). One non-type I and non-type III copper was also found to bind to one molecule of purified bovine factor V by atomic absorption and emission spectroscopy (Mann *et al.*, 1984). Our sequence alignment (not shown) does not indicate a type I or III copper binding site in factor V, but does reveal one putative site for a type II copper-binding at the A1/A3 interface, as in factor VIII.

Third, the activated protein C cleavage site (R562) is situated between β–barrels of A2 domain, a site which is exposed to solvent in our model. This finding is satisfying since proteolytic cleavage sites are often located between structural domains.

Finally, one epitope site, E389-391, implicated by site-directed mutagenesis as major components of inhibitor epitopes in the A2 domain (Ware *et al.*, 1992), is located on the surface of our factor VIII model. Another recently characterized epitope region, 483-509 (Healy *et al.*, 1994), is located in the loop region between β-strands e and f. Part of this region is also exposed to the solvent, as indicated in **Figure III.2d**.

We note that another similar structure could be constructed by making a different choice of amino acid alignments and domain arrangements. Instead of modeling the factor VIII D1, D3 and D5 domains on the NIR D2 domain (and D2, D4, and D6 on NIR D1), the odd-numbered factor VIII domains could be modeled and

spatially superimposed on D1 of NIR and the even-numbered domains on D2. The alternative structure would result in a clockwise, rather than counter-clockwise (as in Figures 1a and 2a), arrangement of domains and would lead to changes in the loops connecting the domains of each A subunit. The type-II copper binding site at the D1/D6 interface and the disulfide linkages are preserved, and the APC and inhibitor sites would remain exposed. We have chosen the former approach because of the stronger sequence similarity and the preservation of the type I copper binding sites in structurally conserved domains, as shown in **figure III.1a.**

One immediate application of this model of the tertiary structure of factor VIII is speculation on how the more than one hundred missense mutations in the coding region of factor VIII can lead to its dysfunction. To date only ten out of 138 missense mutations identified (not including these that affect mRNA splicing) have obvious explanations such as destruction of the thrombin cleavage site and disruption of binding to von Willebrand factor (vWF) (Tuddenham *et al.*, 1994).

Based on our factor VIII model, 33 out of 39 missense mutations associated with severe hemophilia A (Tuddenham *et al.*, 1994) seem likely to destabilize the structure of factor VIII (**Table III.1**), and some mutations are predicted to have multiple debilitating effects. As shown in the solvent accessibility column, nine mutations replace buried hydrophobic residues with either charged or polar amino acids and one substitutes a hydrophilic residue with a hydrophobic residue. 29 mutations change side-chain residue volume by more than 30 $Å^3$, as listed in the third column. Most of these mutations occur in the protein core and are likely to disrupt the tightly packed interior of the molecule. Typically, the energetic penalty for creating a cavity the size of methyl group (33 $Å^3$) is significant, ~1.1 kcal/mole (Eriksson *et al.*, 1992). Mutations involving glycine and proline residues can also have destabilizing effects because these residues have backbone dihedral preferences distinct from the other 18 amino acids. Disruption of factor VIII structure can be

UCSF LIBRARY

Table III.1 Effects of amino acid sustitutions associated with severe hemophilia

Mutation	Solvent Accessibility[1]	Δ Volume (in Å)	Backbone Dihedral (°)	Other Considerations
L7R	**B, bad for + charge**	—	OK	
G22C	B	—	**-86, -160**	
G70D	B	**+58**	OK	
V80D	**B, bad for - charge**	—	OK	
L98R	**B, bad for + charge**	—	OK	
G111R	**B, bad for + charge**	**+102**	**-111,143**	LR
E113D	B	-34	51,-73	LR, **salt bridge with H274**
D116G	B	-111	69,-105	LR, **possible cation binding site with D115**
P146S	**B, bad for OH**	**-61**	OK	
G247E	**B, bad for - charge**	**+46**	**-142,-138**	
G259R	**B, bad for + charge**	**+100**	OK	
R282H	PB	-169	53,79	LR, **salt bridge with R282**
R282L	PB	-100	53,79	LR, **salt bridge with R282**
L308P	B	—	**-122, 126**	
V326L	OK	B	**+44**	
C329Y	PB	+50		**Destroy disulfide bond & creat free SH**
C329R	PB	—		**Destroy disulfide bond & creat free SH**
C329S	PB	-55		**Destroy disulfide bond & creat free SH**
I386S	**B, bad for OH**	**-67**	OK	
E390G	PB	**-137**	OK	?
K425R	PB	—	-6, -63	?
D542G	B	**-89**	OK	LR, **salt bridge with H311 or R541**
I566T	E	+60	144,116	**New glycosolation site**
W585C	PB	**-207**	OK	**Create free SH**
Y586S	B	**-143**	OK	
V634M	B	—	OK	?
E1704K	PB	—	-18, 76	?
G1760D	PB	**+91**	**-86, -82**	LR
M1772T	B	**-109**	150, 94	**New glycosolation site**
S1784Y	B	**+119**	92, 67	
S1784F	B	**+116**	92, 67	
D1846N	B	-36	OK	LR, **salt bridage with H314**
D1846Y	B	+41	OK	LR, **salt bridage with H314**
P1854R	**B, bad for + charge**	—	-138, 165	
R1869I	**E, bad for hydrophobic**	+56 OK	**Salt bridge with E1780**	
E1885K	PB	-36	OK	**Salt bridge with either R1941 or K1943**
N1922S	PB	**-198**	OK	?
N1922D	PB	**-176**	OK	?
R1997W	B	-46	OK	**Salt bridge with Q1799**

[1] B: buried; LR: loop region; PB: partial buried; E: exposed

anticipated in four mutations that change glycines in d-specific conformations to other amino acids or change other amino acids with phi not close to -70° to proline. Ten disrupt existing salt bridges or a possible cation binding site. Five disrupt disulfide bonds or create new free sulfhydryl groups. Free sulfhydryls are disfavored in the extracellular milieu and can complicate the correct folding of proteins. Factor VIII antigen levels were reported as low in three of five tested patients from **Table III.1**. Interestingly our model can not explain the effects of missense mutations (M1772T and V634M) in the two patients with approximately normal levels of factor VIII antigen, indicating that the effects of these mutations is not mediated through gross structural destablization.

The second application of the model could be the development of a second-generation factor VIII. The ready dissociation of the A2 domain, leading to inactivation of the factor VIII, suggests the possibility of genetically engineering a more stable factor VIII by creating type II copper-binding sites between the A1/A2 or A2/A3 domains. Porcine factor VIII, for example, has 10-fold greater activity than human factor VIII because of a lower dissociation rate of the A2 domain from the activation complex (Lollar & Parker, 1991). A more stable factor VIII would improve efficacy and reduce costs and morbidity associated with the treatment of hemophilia A. This might be achieved by introducing histidines at the appropriate positions for the interface between the A1 and A2 domains (F270, Y476 and F536) and/or at those for the A2 and A3 interface (F652, Y1786 and D1840) based on the alignment of factor VIII sequence with AOZ and NIR. The existence of another pathway to inactivate factor VIII, cleavage of the A2 domain by activated protein C (Fay & Walker, 1989), should minimize the thrombosis effect.

A structural model of the A domain of factor VIII will be helpful in analyzing the interaction of factor VIII with other coagulation proteins. For example, vWF stabilizes factor VIII in the conditioned medium by binding factor VIII at several

locations. The main binding sites for vWF are located at the third acidic domain and the C2 domain (Lollar *et al.*, 1988; Norfang & Ezban, 1988; Shima *et al.*, 1993). Particularly, the sulfation of tyr1680 in the third acidic domain appears to be critical in vWF binding (Higuchi *et al.*, 1990; Pittman *et al.*, 1994). Extrapolating from our model, the third acidic and C2 domain could be physically situated in proximity of each other and constitute sites for vWF binding to factor VIII. It is tempting to speculate that five other sulfated tyrosines in factor VIII (346 in the first, 718, 719, 723 in the second and 1664 in the third acidic domain) also participate in vWF binding to some extent.

References

Baker, E. N. (1988). Structure of azurin from Alcaligenes denitrificans refinement at 1.8 A resolution and comparison of the two crystallographically independent molecules. *J. Mol. Biol.* **203**, 1071-95.

Bernstein, F. C., Koetzle, T. F., Williams, G. J., 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. *Eur. J. Biochem.* **80**, 319-24.

Bihoreau, N., Pin, S., Kersabiec, A. M. d., Vidot, F. & Fontaine-Aupart, M. P. (1994). Copper-atom identification in the active and inactive forms of plasma-derived FVIII and recombinant FVIII-delta II. *Eur. J. Biochem.* **222**, 41-48.

Collyer, C. A., Guss, J. M., Sugimura, Y., Yoshizaki, F. & Freeman, H. C. (1990). Crystal structure of plastocyanin from a green alga, Enteromorpha prolifera. *J. Mol. Biol.* **211**, 617-32.

Dunbrack, R. L. & Karplus, M. (1993). Backbone-dependent rotamer library for proteins. Application to side-chain prediction. *J. Mol. Biol.* **230**, 543-74.

Dwulet, F. E. & Putnam, F. W. (1981). Internal duplication and evolution of human ceruloplasmin. *Proc. natn. Acad. Sci. U.S.A.* **78**, 2805-9.

Eaton, D., Rodriguez, H. & Vehar, G. A. (1986). Proteolytic processing of human factor VIII. Correlation of specific cleavages by thrombin, factor Xa, and activated

protein C with activation and inactivation of factor VIII coagulant activity. *Biochemistry* **25**, 505-12.

Eriksson, A. E., Baase, W. A., Zhang, X. J., Heinz, D. W., Blaber, M., Baldwin, E. P. & Matthews, B. W. (1992). Response of a protein structure to cavity-creating mutations and its relation to the hydrophobic effect. *Science* **255**, 178-83.

Fay, P. J. & Smudzin, T. M. (1992). Characterization of the interaction between the A2 subunit and A1/A3-C1-C2 dimer in human factor VIIIa. *Journal of Biological Chemistry* **267**, 13246-50.

Fay, P. J. & Walker, F. J. (1989). Inactivation of human factor VIII by activated protein C: evidence that the factor VIII light chain contains the activated protein C binding site. *Biochim. biophys. Acta.* **994**, 142-8.

Fenderson, F. F., Kumar, S., Adman, E. T., Liu, M. Y., Payne, W. J. & LeGall, J. (1991). Amino acid sequence of nitrite reductase: a copper protein from Achromobacter cycloclastes. *Biochemistry* **30**, 7180-5.

Frieden, E. (1980). Caeruloplasmin: a multi-functional metalloprotein of vertebrate plasma. *Ciba Foundation Symposium* **79**, 93-124.

Godden, J. W., Turley, S., Teller, D. C., Adman, E. T., Liu, M. Y., Payne, W. J. & LeGall, J. (1991). The 2.3 angstrom X-ray structure of nitrite reductase from Achromobacter cycloclastes. *Science* **253**, 438-42.

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

Gregoret, L. M. & Cohen, F. E. (1990). Novel method for the rapid evaluation of packing in protein structures. *J. Mol. Biol.* **211**, 959-74.

Guss, J. M. & Freeman, H. C. (1983). Structure of oxidized poplar plastocyanin at 1.6 A resolution. *J. Mol. Biol.* **169**, 521-63.

Healy, J. F., Lubin, I. M., Nakai, H., Saenko, E. L., Hoyer, L. W., Scandella, D. & Lollar, P. (1994). Residues 484-508 contain a major determinant of the inhibitory epitope in the A2 domain of human factor VIII. *J. Biol. Chem.* **270**, 14505-9.

Higuchi, M., Wong, C., Kochhan, L., Olek, K., Aronis, S., Kasper, C. K., Kazazian, H. H. & Antonarakis, S. E. (1990). Characterization of mutations in the factor VIII gene by direct sequencing of amplified genomic DNA. *Genomics* **6**, 65-71.

Kane, W. H. & Davie, E. W. (1988). Blood coagulation factors V and VIII: structural and functional similarities and their relationship to hemorrhagic and thrombotic disorders. *Blood* **71**, 539-55.

Lollar, P., Hill-Eubanks, D. C. & Parker, C. G. (1988). Association of the factor VIII light chain with von Willebrand factor. *J. Biol. Chem.* **263**, 10451-5.

Lollar, P. & Parker, E. T. (1991). Structural basis for the decreased procoagulant activity of human factor VIII compared to the porcine homolog. *J. Biol. Chem.* **266**, 12481-6.

Mann, K. G., Lawler, C. M., Vehar, G. A. & Church, W. R. (1984). Coagulation Factor V contains copper ion. *J. Biol. Chem.* **259**, 12949-51.

Messerschmidt, A. & Huber, R. (1990). The blue oxidases, ascorbate oxidase, laccase and ceruloplasmin. Modelling and structural relationships. *Eur. J. Biochem.* **187**, 341-52.

Messerschmidt, A., Ladenstein, R., Huber, R., Bolognesi, M., Avigliano, L., Petruzzelli, R., Rossi, A. & Finazzi-Agro, A. (1992). Refined crystal structure of ascorbate oxidase at 1.9 A resolution. *J. Mol. Biol.* **224**, 179-205.

Norfang, O. & Ezban, M. (1988). Generation of active coagulation factor VIII from isolated subunits. *J. Biol. Chem.* **263**, 1115-8.

Petratos, K., Banner, D. W., Beppu, T., Wilson, K. S. & Tsernoglou, D. (1987). The crystal structure of pseudoazurin from Alcaligenes faecalis S-6 determined at 2.9 A resolution. *Febs Letters* **218**, 209-14.

Pittman, D. D., Tomkinson, K. N., Michnick, D., Selighsohn, U. & Kaufman, R. J. (1994). Posttranslational sulfation of factor V is required for efficient thrombin cleavage and activation and for full procoagulant activity. *Biochemistry* **33**, 6952-9.

Pittman, D. D., Wang, J. H. & Kaufman, R. J. (1992). Identification and functional importance of tyrosine sulfate residues within recombinant factor VIII. *Biochemistry* **31**, 3315-25.

Ryden, L. (1982). Model of the active site in the blue oxidases based on the ceruloplasmin-plastocyanin homology. *Proc. natn. Acad. Sci. U.S.A.* **79**, 6767-71.

Shima, M., Scandella, D., Yoshioka, A., Nakai, H., Tanaka, I., Kamisue, S., Terada, S. & Fukui, H. (1993). A factor VIII neutralizing monoclonal antibody and a human inhibitor alloantibody recognizing epitopes in the C2 domain inhibit factor VIII binding to von Willebrand factor and to phosphatidylserine. *Thromb. Haemostasis* **69**, 2406.

Summers, N. L., Carlson, W. D. & Karplus, M. (1987). Analysis of side-chain orientations in homologous proteins. *J. Mol. Biol.* **196**, 175-98.

Takahashi, N., Ortel, T. L. & Putnam, F. W. (1984). Single-chain structure of human ceruloplasmin: the complete amino acid sequence of the whole molecule. *Proc. natn. Acad. Sci. U.S.A.* **81**, 390-4.

Tuddenham, E. G., Schwaab, R., Seehafer, J., Millar, D. S., Gitschier, J., Higuchi, M., Bidichandani, S., Connor, J. M., Hoyer, L. W. & Yoshioka, A. (1994). Haemophilia A: database of nucleotide substitutions, deletions, insertions and rearrangements of the factor VIII gene, second edition. *Nucleic Acids. Res.* **22**, 3511-33.

Vehar, G. A., Keyt, B., Eaton, D., Rodriguez, H., O'Brien, D. P., Rotblat, F., Oppermann, H., Keck, R., Wood, W. I. & Harkins, R. N. (1984). Structure of human factor VIII. *Nature* **312**, 337-342.

UCL LIBRARY

Ware, J., MacDonald, M. J., Lo, M., Graaf, S. d. & Fulcher, C. A. (1992). Epitope mapping of human factor VIII inhibitor antibodies by site-directed mutagenesis of a factor VIII polypeptide. *Blood Coagulation and Fibrinolysis* **3**, 703-16.

Xue, J., Kalafatis, M., Silveira, J. R., Kung, C. & Mann, K. G. (1994). Determination of the disulfide bridges in factor Va heavy chain. *Biochemistry* **33**, 13109-16.

UofL LIBRARY

Chapter 4

Evaluation of Current Techniques for *Ab-Initio* Protein Structure Prediction

UVSF LIBRARY

This chapter has been published in PROTEINS: Structure, Function, and Genetics, Volume 23, pages 431-445, 1995

Introduction

In December, in Asilomar, California, a "Meeting on Critical Assessment of Techniques for Protein Structure Prediction" was held to determine the status of current methods for predicting the three-dimensional structure of proteins. Thirty-five different laboratories attempted, in a blinded fashion, to predict some aspect of the structure of thirty-three different proteins. The structures of these proteins were contemporaneously determined by NMR Spectroscopy or X-ray Crystallography, but were unavailable to the predictors prior to the submission of their predictions. Thus, these represent true or bona fide predictions in the spirit of the work of Schulz and collaborators on adenlyate kinase (Schulz *et al.*, 1974), or Curtis et. al. on Interleukin-4 (Curtis *et al.*, 1991), and not the "retrodictions" of structure that have been called into question by Benner and co-workers (Benner *et al.*, 1992).

The structure predictions fell into three categories: comparative modeling, threading, and *ab-initio* structure prediction. Comparative modeling was defined as structure prediction when the structure of an homologous protein was known (Greer, 1990; Ring & Cohen, 1993; Sali & Blundell, 1993; Summers & Karplus, 1990). Threading predictions were computational attempts to align the sequence of a protein of unknown structure (that lacks clear similarity to another sequence of a protein of known structure) with the side chain environmental preferences dictated by a known protein structure (Bowie *et al.*, 1991; Bryant & Lawrence, 1993; Godzik & Skolnick, 1992; Jones *et al.*, 1992). *Ab-initio* structure predictions attempt to solve the folding problem; given a protein sequence that is unrelated to any

protein of known structure, what is its secondary and tertiary structure? While a great deal of effort has been devoted to this problem, many issues at the secondary structure level and most concerning tertiary structure remain unresolved (Benner & Gerloff, 1991; Chou & Fasman, 1978; Cohen *et al.*, 1986a; Cohen *et al.*, 1979; Crawford *et al.*, 1987; Garnier *et al.*, 1978; King & Sternberg, 1990; Levitt & Warshel, 1975; Rost & Sander, 1994; Russell *et al.*, 1992; Smith & Smith, 1990). Three different laboratories were chosen to evaluate the structure predictions; we evaluated the *ab-initio* predictions.

UVA LIBRARY

Method

Categories of Predictions

The *ab-initio* predictions were divided into four categories: *Class*, *Secondary Structure*, *Fold*, and *Structure*. *Class* prediction was the simplest level of prediction. Predictors evaluated which of the following protein classes the protein most resembled: all α-helix, all β-sheet, α/β or α+β (Cohen *et al.*, 1993; Levitt & Chothia, 1976; Muskal & Kim, 1992; Nishikawa *et al.*, 1983; Sheridan *et al.*, 1985). The *Secondary Structure* category included predictions for each residue of the protein to be in one of three backbone conformations compatible with secondary structure: α-helix, β-sheet, or loop. Predictions of *Fold* described the overall fold or shape of the protein, including many of the common folding motifs originally characterized by J. Richardson (Richardson, 1981) and expanded upon recently by a number of groups (Chothia & Murzin, 1993; Harris *et al.*, 1994; Orengo *et al.*, 1993). Predictions in this category included secondary structure predictions. The final category, *Structure*, was reserved for predictions of the three-dimensional coordinates of the protein. Predictions in this category naturally included specifications for the other three categories. The investigators and the categories in which they made predictions are shown in **Table IV.1**. A short synopsis of their methods is given in **Table IV.2**. Additional information about some of the prediction methodologies can be found in other articles in this issue of the Journal.

Table IV.1 Predictors and Categories

Predictor	Structure	Fold	Secondary Structure	Class
Benner		X	X	
Covell	X			
Garnier			X	
Hubbard		X	X	
Lee	X			
Livingston			X	
Marshall	X			
Meckler		X	X	
Moult	X			
Munson			X	
Osguthorpe	X			
Rose				X
Rost&Sander			X	
Sander	X	X		

UMI

Table IV.2 Synopsis of methods

Investigator	Abbreviation	Method
The ETH Prediction Group: D. Gerloff, G. Chelvanayagam and S.A. Benner	Benner	The prediction method applies automated heuristics to assign surface, interior, active site (tertiary structural information), and parsing residues by analysis of patterns of conservation and variation among homologous protein sequences in light of evolutionary models that interpret amino acid substitutions as the consequence of neutral variation subjected to functional constraints together with adaptive variation that alters the properties of homologous proteins to make them optimally suited to different environments. Secondary structural elements are assigned from patterns in the tertiary structural information (Benner *et al.*, 1994).
B.K. Lee and N. Kurochkina	Lee	For a polypeptide chain, a biased Monte Carlo search was applied for the dihedral angles of the main chain phi and psi and side chain dihedral angels chi. Conformational space was reduced into a small number of allowed regions in Ramachandran phi and psi map (Kang *et al.*, 1993). Weighted sum of hydrophobic energy based on pairwise surface area sum(Kurochkina & Lee, 1995, in press) and hydrogen bond energy calculated as electrostatic Coulomb sum was used to estimate the energy of the structure.

UVIC LIBRARY

S.G. Galaktionov and G.R. Marshall	Marshall	The secondary structure of the protein is predicted using a consensus of three methods implemented in SYBYL 5.5. Next an algorithm was used to predict coordination number vectors for the amino acid residues (Rodionov & Galaktionov, 1992). Then the residue-residue contact matrix was predicted using an iterative procedure to improve heuristic gain function. Finally, the spatial structure was reconstructed (Galaktionov & Marshall, 1994).
J.T. Pedersen and J. Moult	Moult	A torsion space representation of a protein is used with an all atom force field (Avbelj & Moult, 1995) together with a genetic algorithm (Pedersen & Moult, Document in preperation) and a Monte-Carlo algorithm (Avbelj & Moult, 1995, in press) to predict the structure of small proteins.
D.J. Osguthorpe	Osguthorpe	A simplified model of protein structure with potentials developed to reproduce the physical behavior of atoms rather than protein statistics derived from the database. The potentials are being continuously improved to reproduce protein-like structures.
T. Hubbard and J. Park	Hubbard	Automatic alignment of sequences using the PHD server (Rost & Sander, 1994) followed by addition of more sequences and hand alignment. These alignments were then submitted to the PHD neural network in Heidelberg. Fold prediction was aided by a strand pairing algorithm (Hubbard, 1994).
L. Holm, B. Rost, P. Bork and C. Sander	Sander	Secondary structure was predicted for all proteins using the neural network method that uses sequence profiles as input (Rost & Sander, 1994).

UCSF LIBRARY

B. Rost and C. Sander	Rost&Sander	The secondary structure elements were then assembled into three-dimensional structures.
G. Livingston and H.B. Nicholas	Livingston	Case based learning approach. Various 22 amino acid segments are compared to the protein to be predicted, if the sequence matching score exceeds a threshold, the structure of the 22 amino acid segment is used as evidence to predict the secondary structure (Leng, 1994; Leng *et al.*, 1994).
J. Garnier and J.M. Levin	Garnier-SIMPA	SIMPA (SIMilarity Peptide Analysis) program is based on sequence similarity between a stretch of amino acids (17 amino acid long) of the test sequence and the sequences in a data base of protein structures. Q3 of 86% when a homologous protein structure is present, otherwise 63-65% (Levin & Garnier, 1988). When homologous sequences are known, it can be associated with the CONSENSUS program (Levin *et al.*, 1993) to yield an accuracy of 68-69%.
J. Garnier and V. DiFrancesco	Garnier-COMBINE	The COMBINE method is an expert system amalgamation of three secondary structure prediction algorithms: GOR III, SIMPA and Bit Pattern (Biou *et al.*, 1988). It can be associated with multiple sequence alignments (CONSENSUS) to yield an accuracy of 69-70%(DiFrancesco *et al.*, 1995).

UNT LIBRARY

P. J. Munson and V. DiFrancesco	Munson	Two different multiple sequence methods: QL(quadratic logistic), Profile-QL. The QL method is a calibrated logistic model for a three state prediction using the maximum likelihood principle (Munson *et al.*, 1994). The profile method combines this method with multiple sequence alignment information (DiFrancesco *et al.*, 1995). The expected accuracy for Profile-QL is 67% to 69% measured in two separate crossvalidated tests.
R.G. Idlis and L.B. Mekler	Mekler	Prediction of specific contacts between amino acid residues of the protein molecule being in the intermediate conformation, the so called "molten globule". These contacts are supposed to be determined by the specific binding of amino acid residues encoded by a codon and its anticodon. The folding of an amino acid sequence into the "molten globule" is a step-by-step co-translational process of the formation and reorganization of these code bonds. An additional sterochemical code is supposed to determine the first order phase transition that underlies protein activity. It is supposed that the two conformations of a protein molecule have a similar topology of the backbones by the entirely different systerms of hydrogen bonds and Van der Waals interatomic contacts (Mekler & Idlis, 1993).
D. Covell	Covell	Simulated annealing methods are applied to a simple cubic lattice alpha-carbon model of a protein. Each amino acid occupies only one lattice site. Several simulations of greater than 100,000 steps are carried out to determine the consensus configuration of the protein (Covell, 1992; Covell, 1994).

R. Srinivasan and G. Rose	Rose	Not specified at the time of submission.

UNDER

Evaluation of Predictions

The success or failure of class prediction was decided by visually assigning a class to each protein and comparing to the predicted class (Levitt & Chothia, 1976).

Secondary structure predictions were evaluated by comparing the predicted with the experimentally determined secondary structure. The percentage correct score in a three-state system (Q3: α-helix, β-sheet, or loop) was used (Schulz & Schirmer, 1979). The secondary structure of the experimentally determined structure was calculated with the program DSSP (Kabsch & Sander, 1983) which assigns secondary structure by examining hydrogen bonding patterns in the context of backbone dihedral angle preferences. The secondary structure predictions were further evaluated by subcategorizing the incorrect predictions into three categories: OVER, UNDER, and WRONG. OVER was defined as predicting an α-helix or β-sheet when the protein formed an aperiodic or loop structure in reality. UNDER was defined as predicting a loop conformation when the residue adopted an α-helical or β-sheet geometry. WRONG was defined as predicting an α-helix when the amino acid was in a β-sheet or vice-versa. While one can imagine molecular dynamics simulations or other optimization methods correcting UNDER and OVER prediction, WRONG predictions are likely to be extremely difficult to recover from. For each protein, the secondary structure prediction methods were compared to the GOR (Garnier *et al.*, 1978) method as an historical standard.

The overall fold of the protein was evaluated qualitatively, from a visual comparison of the experimentally determined structure with the predicted description.

The precise structure of the protein was evaluated by the root mean square deviation (r.m.s.) between equivalent alpha carbons in the predicted and experimental structures. This calculated value was compared to the r.m.s. value expected for random compact structures (Cohen & Sternberg, 1980a).

The structure of the protein was also evaluated using a recently developed method that minimizes the area of a "soap film" that would join the predicted and experimentally determined poly-peptide backbone. Benchmarks for this type of comparison have been developed to help to assess if any of the predicted models have captured features of the chain topology and fold (Falicov & Cohen, 1996).

Results and Discussion

All the predictors taking part in this contest should be congratulated. Many of the structure predictions made were completed under less than ideal conditions. Some prediction methods, which typically require six months to apply in their entirety, were made in one. Some of the prediction strategies applied remain in a developmental phase and so these predictions should be regarded as work in progress. For this reason, we are stressing the promising results from the meeting, while still noting all of the results.

The results of the *Structure* predictions are shown in **Table IV.3**. The *Fold* predictions are shown in **Table IV.4**; the *Secondary Structure* predictions are shown in **Figure IV.1**. The *Class* predictions are shown in **Table IV.5**.

Overview

The main issue in this section of the conference was whether or not it is presently possible to predict *ab-initio* the tertiary structure of proteins. Two different approaches were used to predict tertiary structures. The first was Primary → Secondary → Tertiary which involved predicting the secondary structure of the protein from the amino acid sequence(s) (Primary → Secondary) and then assembling a tertiary structure from the secondary structure elements (Secondary → Tertiary). The second method involved going straight from the sequence(s) to the tertiary structure. These two techniques met with varying success.

Table IV.3 Evaluation of protein structure prediction

* The standard deviation associated with the average Random r.m.s. scores is +/- 1.4 Å(Cohen & Sternberg, 1980a).

‡ The r.m.s. could not be determined due to a reversal in chain tracing

† A soap bubble value of <0.35 is somewhat accurate < 0.3 is adequate < 0.25 is good < 0.2 is very good(Falicov & Cohen, 1996).

Protein	**Length (residues)**	**Random R.M.S. *(Å)**	**Predictor**	**R.M.S. (Å)**	**Soap Bubble**	**Energy**
Membrane Binding domain for the C2 domain of human coagulation factor VIII	22	10.3	Moult	4.4	0.15	-46.1
				8.8	0.29	-45.1
				9.1	0.35	-41.2
			Lee	4.4	0.14	-236
				7.7	0.23	-212
Subtilisin Propiece	71	12.6	Marshall	11.4	0.35	
Subtilisin Propiece segment	16	10.0	Moult	10.2	0.43	
Domain 3 of Staufen	68	12.4	Marshall	13.7	0.36	
			Osguthorpe	19.6	0.33	

				21.3	0.31	
				12.9	0.35	
Chymotrypsin / Elastase Inhibitor-1	63	12.2	Covell	7.3	0.32	
6-phospho-beta-D-galactosidase	454	30.4	Sander	‡	0.26	

Table IV.4 Evaluation of protein fold prediction

Protein	Actual Fold	Predictor	Prediction
6-phospho-beta-D-galactosidase	α/β barrel	Benner	α/β barrel
		Sander	α/β barrel
Xylanase	α/β barrel	Sander	α/β structure with one or more β-sheets rather than a closed barrel
Biphenyl-2,3-Diol 1,2-Dioxygenase	Two symmetrical regions split into two regions of E-H-E-E-E	Hubbard	Two symmetrical regions split into two regions of E-H-E-E-E-E
Membrane Binding domain for the C2 domain of human coagulation factor VIII	α-helix with a twist on the end	Moult	α-helix with a twist on the end
			disordered β-structure
			short helix packed against a strand/coil
		Lee	α-helix with a twist on the end
			α-helix with a β-strand pair at the end
Chorismate Mutase	All α-helical dimer with a coiled coil along the N-terminal helix	Hubbard	All α-helical dimer with a coiled coil along the N-terminal helix

Domain 3 of Staufen	Two α-helices packed against the same face of a three stranded β-sheet	Hubbard	Two α-helices packed against the same face of a two stranded β-sheet
		Mekler	Two α-helices packed against opposite sides of a two stranded beta sheet
		Osguthorpe	Disordered β structure
			N-terminal α-helix and disordered coil
			Compact disordered coil
		Marshall	Two α-helices packed against opposite sides of a two stranded beta sheet
Chymotrypsin / Elastase Inhibitor-1	Coiled Structure with five disulfide bonds	Covell	Coiled Structure with five disulfide bonds.
Replication Terminator Protein	α+β Leucine Zipper Dimer	Hubbard	All α–helical protein making a Leucine Zipper dimer
		Mekler	α + β dimer differing in placement of the secondary structure regions, resulting in different overall fold.

Synaptotagmin I C2	β-Sandwich	Benner	Pleckstrin like seven β-strands plus one α-helix
		Hubbard	Pleckstrin like seven β-strands plus one α-helix
Subtilisin Propiece	Three strand β-sheet packed against two helices	Marshall	One stranded β-sheet with helices one either side
Subtilisin Propiece segment	extended	Moult	β-hairpin

Table IV.5 Evaluation of protein class prediction

Protein	**Class**	**Predictor**	**Prediction**
Chorismate Mutase	All α-helical	Rose	All α-helical
Synaptotagmin I C2	All β-sheet	Rose	All β-sheet

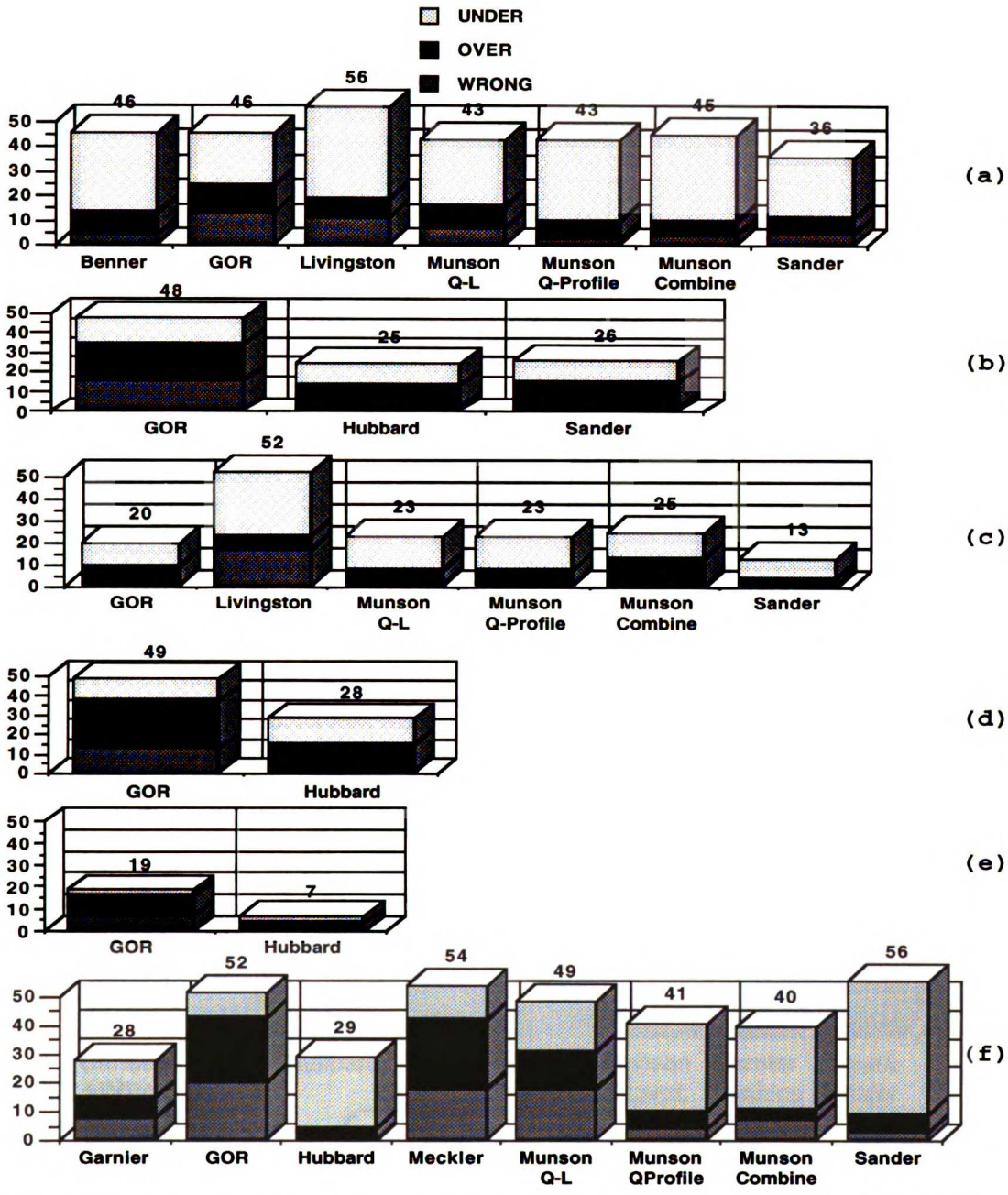
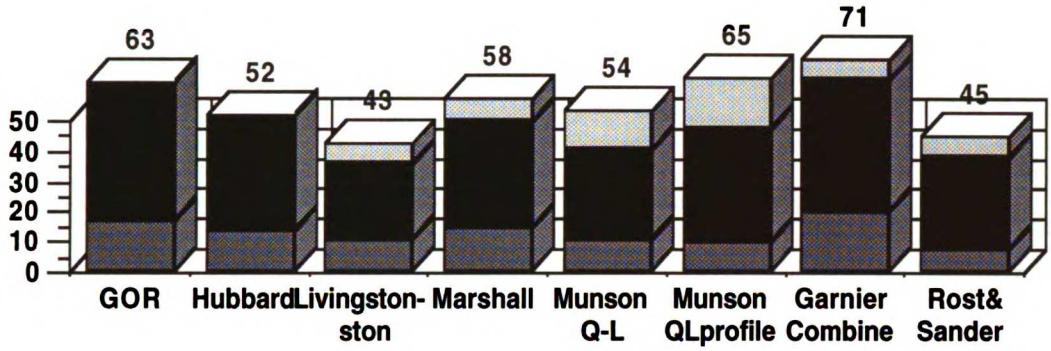
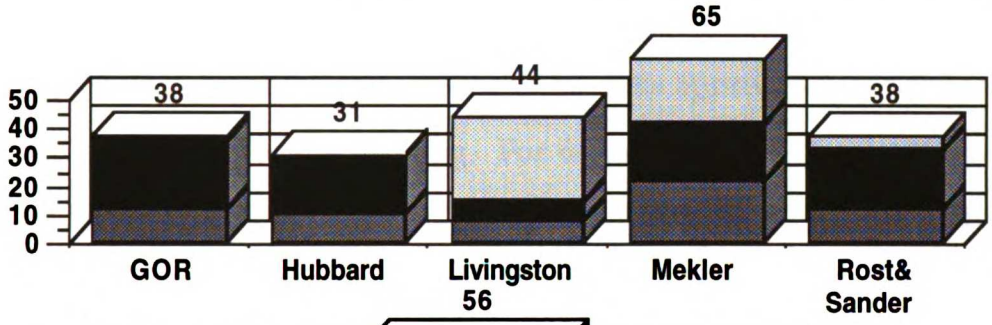


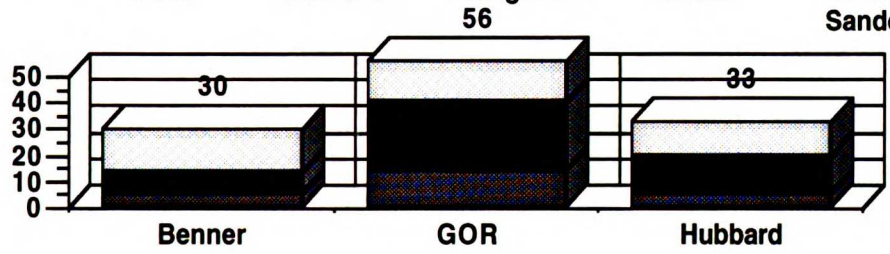
Figure IV.1 Secondary structure predictions for proteins broken down into three categories: OVER, UNDER, WRONG (see text for definitions). The percentage score above the histogram is the total percentage incorrect predictions. (a) 6-phospho-beta-D-galactosidase (b) Xylanase (c) Mystery (d) Biphenyl-2,3-Diol 1,2-Dioxygenase (e) Chorismate Mutase (f) Domain 3 of Staufen (g) Subtilisin Propiece (h) Replication Terminator Protein (i) Synaptotagmin I C2 (j) prokaryotic ribosomal protein l14 (k) Pyruvate phosphate (l) Klebsiella aerogenes Urease: Beta and Gamma subunits. (*g,h,i,j,k and l are on the following page*)


63
52
43
58
54
65
71
45
50
40
30
20
10
0
GOR
Hubbard
Livingston-ston
Marshall
Munson Q-L
Munson QLprofile
Garnier Combine
Rost& Sander

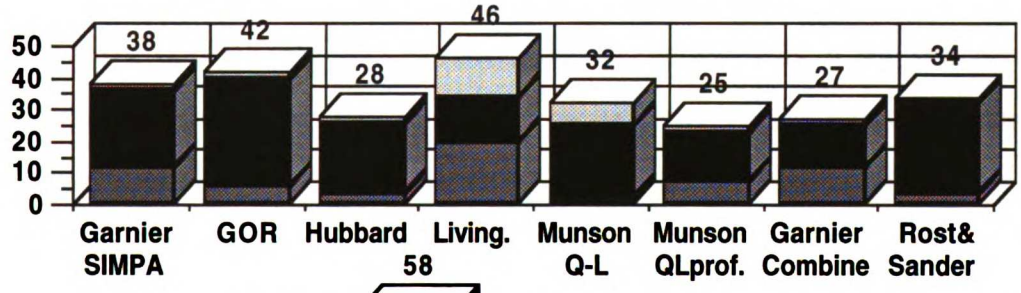
(g)


38
31
44
65
38
50
40
30
20
10
0
GOR
Hubbard
Livingston
Mekler
Rost& Sander

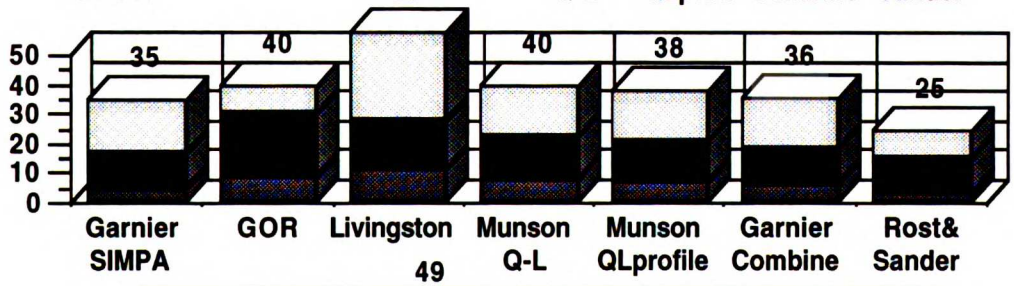
(h)


56
30
33
50
40
30
20
10
0
Benner
GOR
Hubbard

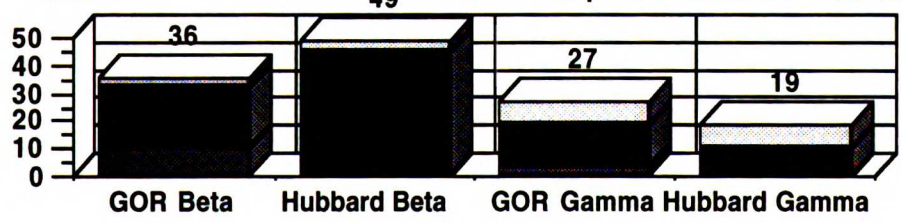
(i)


38
42
28
46
32
25
27
34
50
40
30
20
10
0
Garnier SIMPA
GOR
Hubbard
Living. 58
Munson Q-L
Munson QLprof.
Garnier Combine
Rost& Sander

(j)


35
40
40
38
36
25
50
40
30
20
10
0
Garnier SIMPA
GOR
Livingston
Munson Q-L
Munson QLprofile
Garnier Combine
Rost& Sander

(k)


36
49
27
19
50
40
30
20
10
0
GOR Beta
Hubbard Beta
GOR Gamma
Hubbard Gamma

(l)

Primary → Secondary → Tertiary

Primary → Secondary

Q3 Can be a Misleading Measure of Prediction Accuracy

The first step in predicting the structure of a protein following the Primary → Secondary → Tertiary approach is predicting the secondary structure. Secondary structure predictions are traditionally evaluated using a three state percentage correct score, Q3. This approach was used to evaluate each prediction for the meeting. For 6-phospho-beta-D-galactosidase, Benner and Sander correctly predicted this protein to be an α/β barrel. However, their secondary structure prediction accuracy differed considerable, 67% for Benner compared to 75.4% for Sander. Benner's level of accuracy was more similar to that of the GOR algorithm (Garnier *et al.*, 1978), which correctly predicted 62% of the secondary structure of this protein (see **Figure IV.2**). With these modest per-residue prediction scores, how was Benner able to predict the correct fold? The answer lies in the analysis of the secondary structure prediction shown in **Figure IV.1a**. This figure demonstrates that Benner, Rost&Sander, Munson, and Garnier-COMBINE all had an exceptionally small number of "WRONG" predictions. From the viewpoint of fold prediction, the correct assignment of amino acids comprising the structural core of the protein is more important than the conformational assignments for amino acids that form the end of secondary structure elements and loop regions. For example, the structure of 6-phospho-beta-D-galactosidase has a large excursion from the standard α/β barrel fold (see **Figure IV.3**). This structure was missed by both Rost&Sander and Benner, and counts in the UNDER category. In addition, the exact beginnings and ends of secondary structure are less

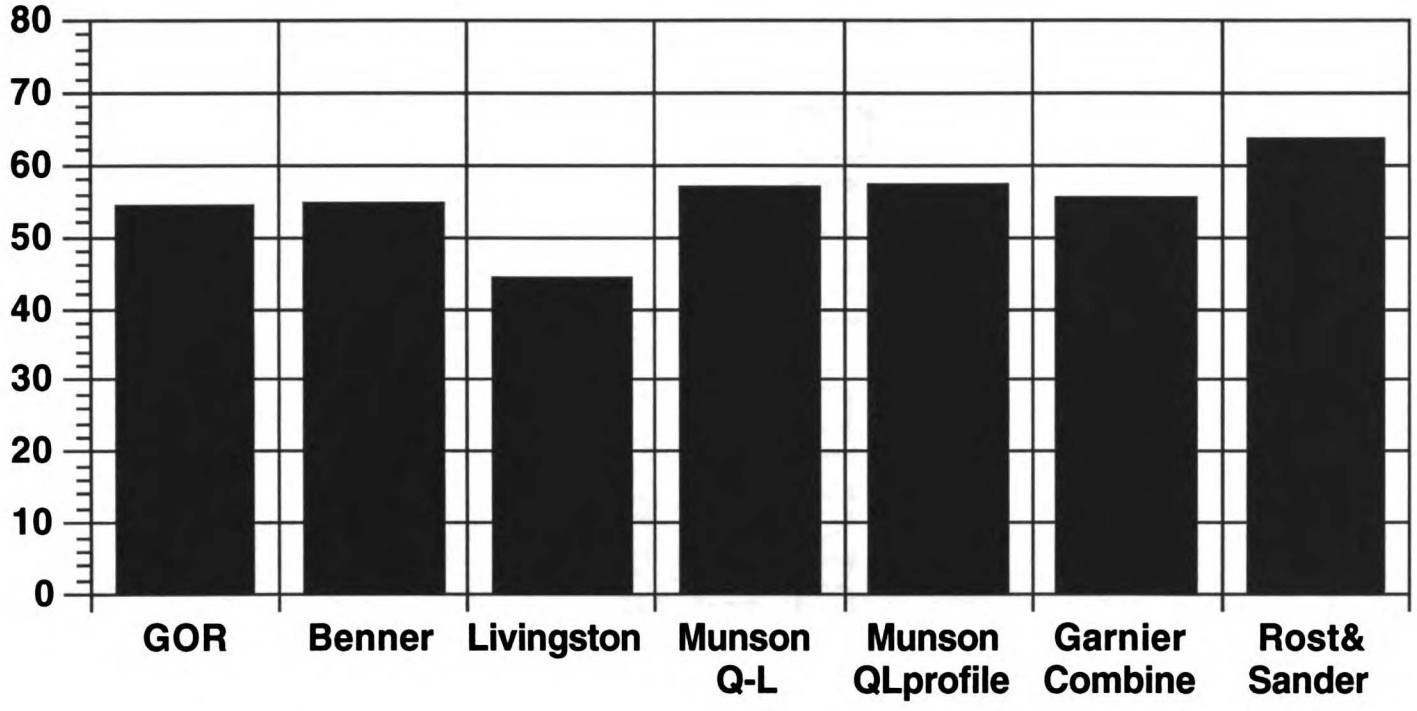


Figure IV.2. Percent correct secondary structure of 6-phospho-beta-D-galactosidase in a three state system (α-helix, β-sheet, or other). The secondary structure of the experimentally determined structure was calculated by DSSP (Kabsch & Sander, 1983) .

Figure IV.3. 6-phospho-beta-D-galactosidase. Picture generated by midas Ribbonjr (Ferrin *et al.*, 1988). *b*-strands are in red, *a*-helices are in green, and the rest of the chain is in purple. Secondary structure calculated by DSSP (Kabsch & Sander, 1983).

important for fold prediction. Errors in these places result in UNDER or OVER prediction. **Figure IV.4** shows the parts of 6-phospho-beta-D-galactosidase predicted correctly and as WRONG for both the Benner prediction and standard GOR predictions. It can be seen how a correct fold prediction was possible for Benner, but would have been unlikely if a fold prediction was made from the GOR secondary structure prediction.

Benefits of Multiple Sequence Alignments

The main difference between the GOR algorithm and the secondary structure prediction methods demonstrated in this meeting is the exploitation of the structural information implicit in multiple sequence alignments. Since each sequence in the alignment codes for approximately the same structure, the secondary structure elements for each of these proteins should co-localize. This redundancy of information allows the central portion of most secondary structure regions to be assigned correctly (very few WRONG assignments). The ends of secondary structure regions vary between the sequences, and as expected OVER and UNDER prediction remains common in methods based on aligned sequence information.

It is thus not surprising that for Xylanase, a more regular α/β barrel than 6-phospho-beta-D-galactosidase, the secondary structure prediction accuracy for the multiple sequence methods improved (**Figure IV.1b**), while the GOR method maintained the traditional level of accuracy.

As another example, each of T. Hubbard's predictions was analyzed, and compared to the GOR score for the same proteins. The average percent

Figure IV.4. 6-phospho-beta-D-galactosidase. Correct α–helix predictions are shown in purple, correct β-strand predictions are shown in green, and WRONG (α–helix predicted for β-sheet region or β-sheet region predicted for α-helix region) predictions in red. The prediction in (*a*) was done by Benner. The prediction in (*b*) was done with the 1977 version of the GOR algorithm. (*figure b is on the following page*)


(b)

correct secondary structure per protein was 68.6% for Hubbard and 58.3% for GOR. However, the difference was much more pronounced with respect to the WRONG predictions. Hubbard had only 2.3% WRONG predictions while the GOR method produced WRONG predictions for 10.2% of the residues. The total errors for the OVER and UNDER predictions was 29.0% for Hubbard and 31.5%% for GOR - almost identical.

When the multiple sequence alignment is limited in sequence number or covers a narrow phylogeny, the quality of the secondary structure prediction suffered. For example, the Replication Terminator Protein had only two homologous sequences in the sequence data banks. **Table IV.6** shows the number of sequences present in a family of aligned homologous sequences for each protein. **Figure IV.1h** demonstrates the poor results. This effected the overall fold predictions as shown in **Table IV.4;** neither of the fold predictions were correct.

Synaptotagmin I C2 presented a different alignment problem. **Table IV.6** shows that Synaptotagmin I C2 had 40 sequences in its multiple sequence alignment. However, the C terminal end of the protein showed a large amount of sequence divergence. Both Hubbard and Benner correctly predicted the first six strands of Synaptotagmin I C2, but the C terminal strand was mispredicted to be a helix by both labs.

Prokaryotic ribosomal protein L14 demonstrated that even with the highly unusual structure shown in **Figure IV.5**, accurate secondary structure assignments can be made with a sufficiently large family of aligned sequences, 25 sequences in this case. **Figure IV.1j** shows the secondary structure prediction.

Table IV.6

Number of aligned sequences for predicted proteins*

*Each of the sequences in the alignment was considered to be homologous to the probe sequence when the sequence identity was ≥ 30%.

Protein	Sequences*
6-phospho-beta-D-galactosidase	16
Mystery	2
Xylanase	13
Biphenyl-2,3-Diol 1,2-Dioxygenase	16
Membrane Binding domain for the C2 domain of human coagulation factor VIII	1
Pyruvate Phosphate Dikinase	6
Chorismate Mutase	7
Domain 3 of Staufen	8
Klebsiella Aerogenes Urease beta	9
Klebsiella Aerogenes Urease gamma	11
Chymotrypsin / Elastase Inhibitor-1	1
Replication Terminator Protein	3
Synaptotagmin I C2	40
Prokaryotic Ribosomal Protein l14	25
Subtilisin Propiece	12

UMI THESIS

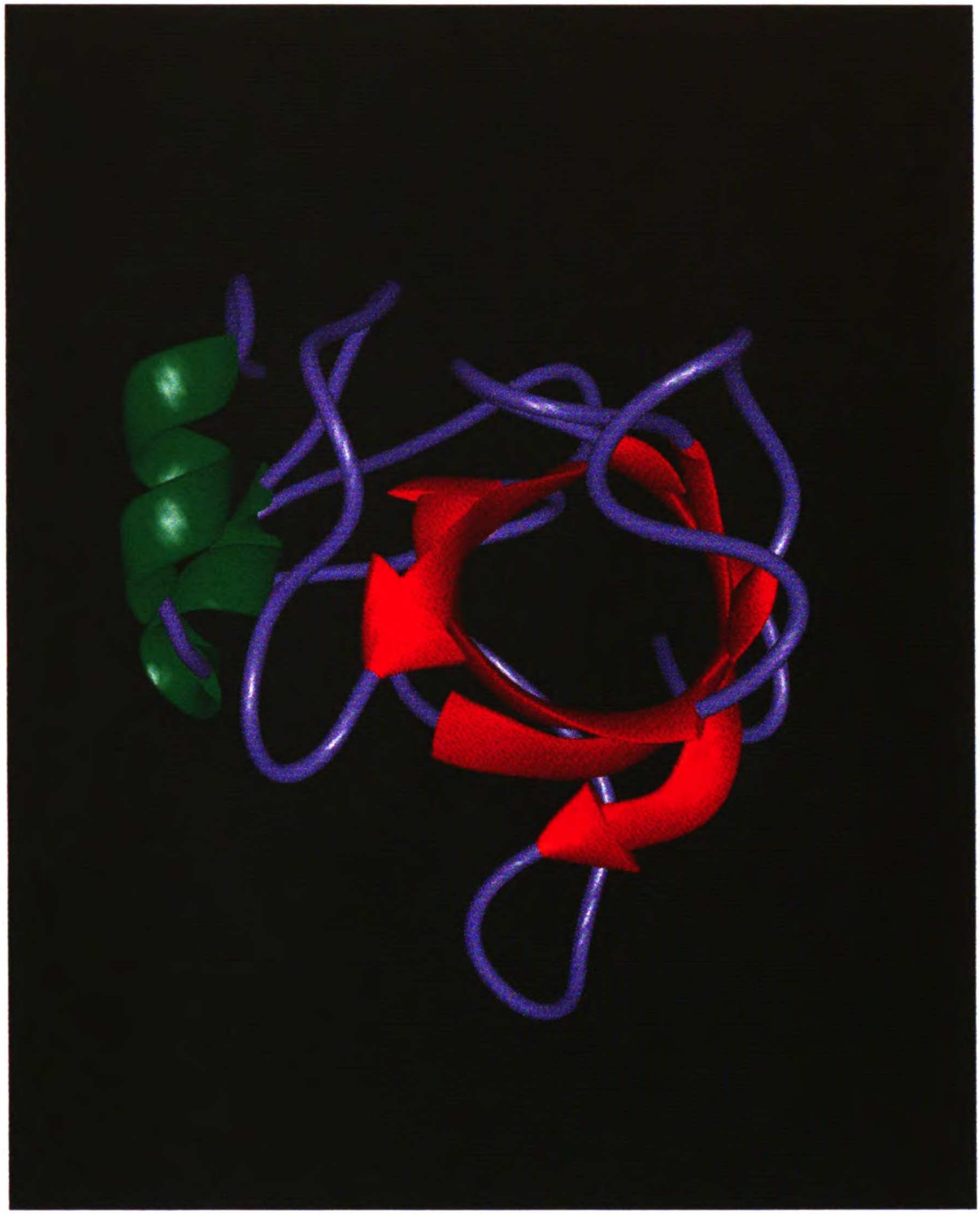


Figure IV.5. Prokaryotic ribosomal protein l14. Picture generated by midas Ribbonjr (Ferrin *et al.*, 1988). β-strands are in red, α-helices are in green, and the rest of the chain is in purple. Secondary structure calculated by DSSP (Kabsch & Sander, 1983).

Is human intervention superior to totally automated approaches?

A point which has been argued in the literature is whether or not human intervention is superior to totally automatic methods in secondary structure prediction (Benner & Gerloff, 1993; Robson & Garnier, 1993). For a group of proteins, Hubbard and Rost&Sander both used the same secondary structure prediction algorithm: PHD (Rost & Sander, 1994). However, Hubbard aligned the sequences automatically and optimized them by hand; Rost&Sander's alignment method was totally automated. **Figure IV.6** indicates that Hubbard's hand alignment improved the ability of PHD to accurately predict secondary structure.

Another example of Man vs. Machine is shown by the predictions made by Benner and coworkers. Although much of Benner's technique is now automated, there is still a large human element in the structure predictions that their lab performed. **Figure IV.7a and IV.7b** demonstrate that their method has a similar level of effectiveness for secondary structure prediction as many of the automated methods.

What Have We Learned About Secondary Structure Prediction?

The Mystery protein demonstrates the correspondence between the information we use to predict secondary structure and that used to design proteins. Mystery was a designed TIM barrel called RORO. It was designed by the first EMBO protein design course, improved by Chris Sander and Gert Vriend(Sander & Vriend, 1992), and produced by Steve Emery (Emery & Fritz, 1994). The extremely accurate secondary structure predictions shown in **Figure IV.1c** indicate that the rules used to design this TIM barrel strongly

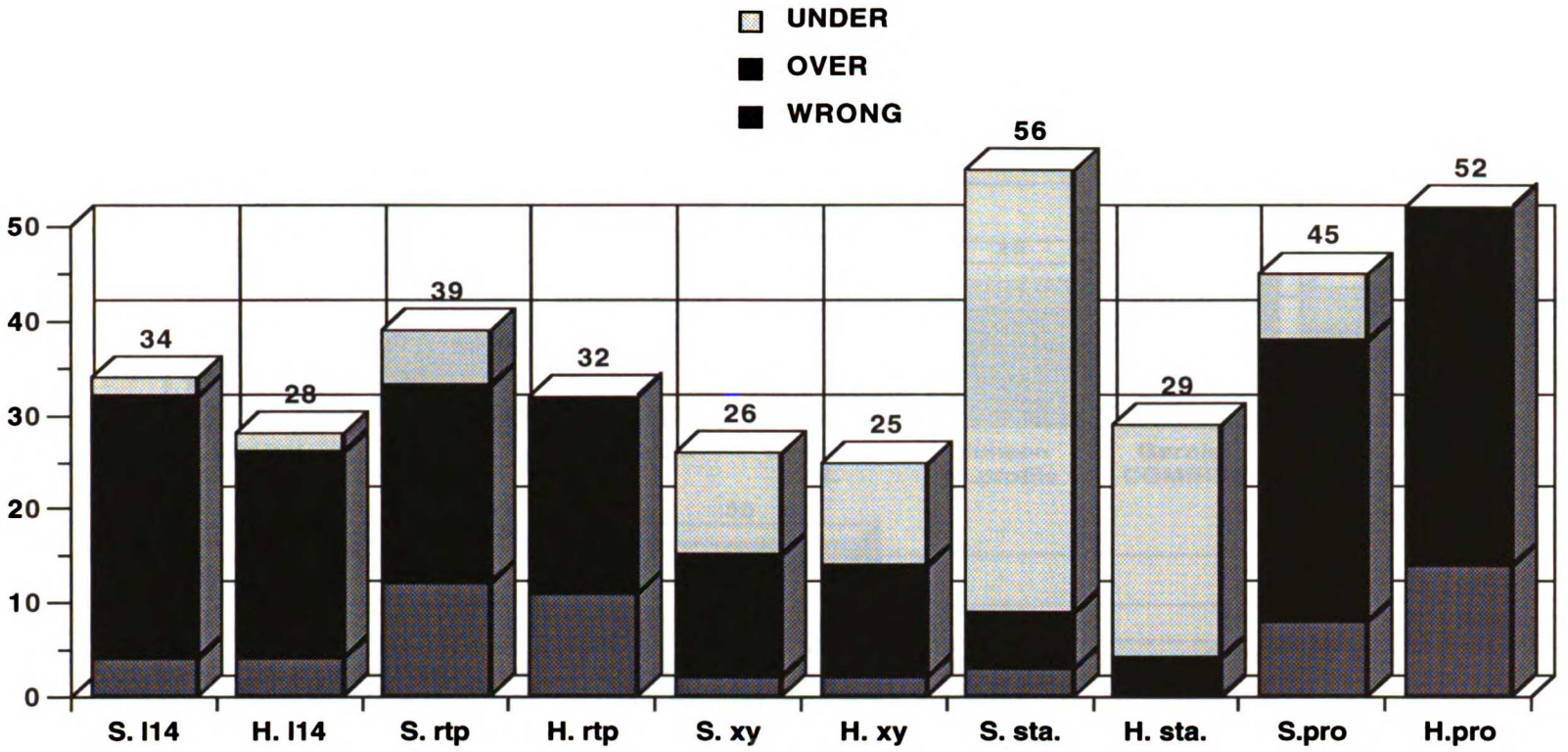


Figure IV.6. Comparison of automated sequence alignment versus automated plus hand alignment. Secondary structure predictions were broken down into three categories OVER, UNDER, WRONG. The percentage score above the histogram is the total percentage incorrect predictions. Rost&Sander and Hubbard are abbreviated S and H. Proteins evaluated include Ribosomal protein l14 (l14), Replication Terminator Protein (rtp), Xylanase (Xy), Domain 3 of Staufen (Sta), and Subtilisin propiece (Pro). The secondary structure of each of these proteins was predicted by both investigators. Hubbard used automated alignments supplemented by hand alignment, Rost&Sander used automated alignments. They both submitted the alignments to Rost&Sander's PHD server (Rost & Sander, 1994) .

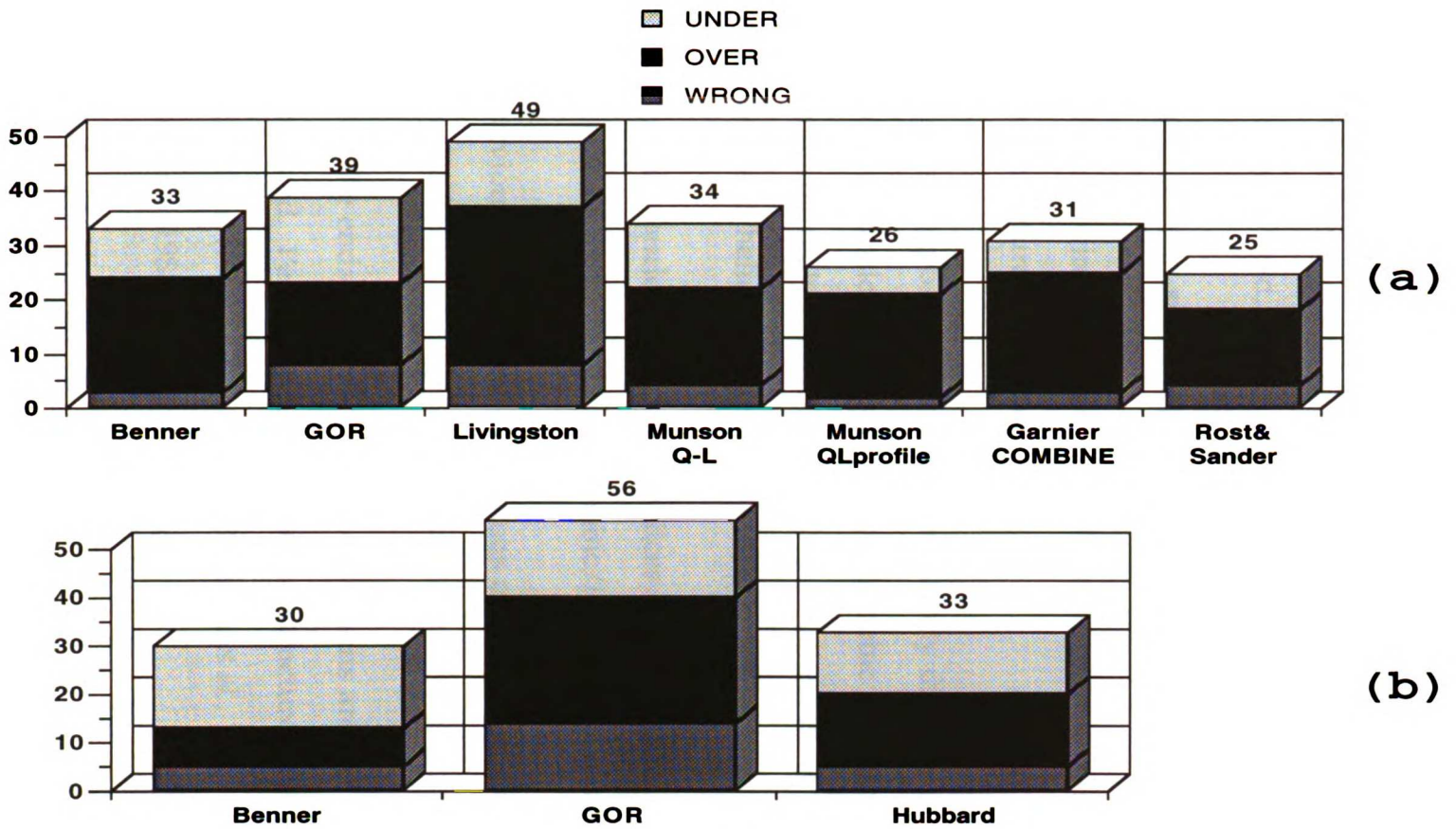


Figure IV.7. Comparisons of the Benner prediction of (*a*) 6-phospho-beta-D-galactosidase with several other investigators and of (*b*) Synaptotagmin I C2 with Hubbard. Secondary structure predictions were broken down into three categories OVER, UNDER, WRONG. The percentage score above the histogram is the total percentage incorrect predictions. This is a comparison between human guided structure prediction (Benner) and automated approaches (Hubbard and others).

resemble the rules used to predict its structure. These are not however, the rules nature uses for folding proteins; RORO showed approximately the correct helical content but almost no beta sheet by CD and NMR spectroscopy(Schmid, 1994).

Difficult protein substructures

The following are some examples of specific errors that occurred in secondary structure prediction.

- Completely exposed helices are consistently difficult to predict due to the lack of the classical hydrophilic, hydrophobic repeating pattern of the more common partially buried helices. For instance, most groups missed one of the exposed helices in Ribosomal Protein L14.

- Completely buried strands and helices are also difficult to predict due to their lack of a repeating hydrophobic/hydrophilic pattern.

- As previously mentioned, excursions of secondary structure not present in the entire family of aligned sequences are difficult to predict.

- Finally, the ends of secondary structure units are still frequently misassigned. Perhaps work on capping structures will serve to address this problem (Harper & Rose, 1993; Levin *et al.*, 1993; Richardson & Richardson, 1988; Zhou *et al.*, 1994).

Secondary → Tertiary

Approaches to Secondary Structure Assembly

One approach to assembling the structure of a protein is to attempt to combine the secondary structure units of the protein in every plausible way, and evaluate which assembly is most likely to be correct (Cohen *et al.*, 1979; Cohen *et al.*, 1980; Cohen *et al.*, 1982; Ptitsyn & Rashin, 1975). Benner attempted to assemble Synaptotagmin I C2 by this combinatorial approach. Unfortunately, they mispredicted a strand for a helix. Even so, the correct overall fold was present in their list of plausible folds. This fold was rejected in the evaluation stage (**Table IV.4**).

The task of assembling secondary structure units is simplified when the secondary structure exhibits a pattern seen before. The overall fold of 6-phospho-beta-D-galactosidase was determined largely because the repeated α-β pattern was familiar to the investigators (**Table IV.4**). Sander was also able to propose a structure with coordinates (**Table IV.3**). Even when the secondary structure exhibits a familiar pattern, mistakes are possible. Sander rejected an α/β barrel in favor of an alternative α/β structure with one or more sheets rather than a closed barrel (**Table IV.4**).

Assembly of the secondary structure units into a fold is also aided by local folding motifs such as the leucine zipper(Landschulz *et al.*, 1988; O'Shea *et al.*, 1989). This motif is formed when two helices pack against one another with leucines at the interface (See **Figure IV.8**). The fingerprint of this motif is a leucine repeated every seventh amino acid. Hubbard was able to recognize this motif and correctly predict a leucine zipper in Chorismate Mutase. Since only one leucine zipper helix was shown, Hubbard correctly predicted that Chorismate Mutase was "an all helical dimer with a coiled coil along the N-terminal helix." Another example of the leucine zipper motif was seen in the

Replication Terminator Protein. In this case again the presence of a dimer was identified.

When the protein to be predicted has an unusual fold, it is more difficult to assemble the secondary structure units. Two of the proteins that several labs made fold predictions for had an unusual fold: Domain 3 of Staufen and Subtilisin Propiece. They both had a beta sheet sandwiched against a pair of helices. This can be contrasted with the more common motif of helices covering both sides of a β-sheet (Cohen *et al.*, 1982).

Marshall and Mekler's groups both predicted Domain 3 of Staufen to have the two helices on opposite sides of the sheet. Hubbard's group correctly predicted that the helices would lie on the same side of the sheet. The predicted folds with coordinates are shown in **Figure IV.9**. The r.m.s. and soap bubble values are shown in **Table IV.3**.

Another unusual motif was the Subtilisin Propiece, although its structure is surprisingly similar to that of Domain 3 of Staufen. As shown in **Table IV.3**, Marshall predicted the overall fold with an r.m.s. error of 11.4 Å. Moult predicted the conformation of residues 7-22 with an r.m.s. error of 10.2 Å. The unusual folds seem to have effected the secondary structure predictions as well. As seen in **Figures IV.1f** and **IV.1g**, the secondary structure predictions are frequently in error. It is possible that common folds have improved secondary structure prediction accuracy due to the presence of similar structures in the databases used to derive prediction parameters.

Figure IV.8. Chorismate Mutase. Leucines are yellow. One subunit is in white and the other in red. Secondary Structure evaluated using DSSP. Picture generated by Midas Ribbonjr (Ferrin *et al.*, 1988) .

Figure IV.9. Domain 3 of Staufen. (*a*) experimentally determined X-ray crystal structure (*b*) Osguthorpe prediction (*c*) Marshall prediction. Structures generated by Midas Ribbonjr (Ferrin *et al.*, 1988) . The N-terminus is red, the C-terminus is blue, and the middle of the sequence is green.

The high symmetry of the TIM barrels clearly aided prediction of the overall fold of proteins in this study as well as in previous efforts (Crawford *et al.*, 1987). This effect was also seen with Biphenyl-2,3-Diol 1,2-Dioxygenase whose overall fold was predicted by Hubbard to be "two symmetrical regions, each split into two E-H-E-E-E-E regions," where E is an extended β-strand and H is an α-helix. In reality each region is E-H-E-E-E. The secondary structure prediction is shown in **Figure IV.1d**. Gene duplication and other evolutionary mechanisms frequently lead to proteins with substantial internal symmetry. Clearly, this can be put to advantage in prediction efforts. When recognized, symmetry elements can serve to multiply the amount of homologous sequence information and frequently extends the phylogenetic separation between structurally related elements. This was observed with the internal two fold identified by Hubbard in Biphenyl-2,3-Diol 1,2-Dioxygenase, as well as the implicit four fold in four-helix bundles or eight fold symmetry in α/β barrels.

Are We Really Just Threading?

Threading matches a protein sequence with known protein structures. The two correct α/β barrel prediction were essentially matching the patterns of secondary structure of the unknown protein to that of known proteins. Benner's misprediction of Synaptotagmin C2 was partly due to mispredicted secondary structure. The secondary structure pattern they predicted matched the pattern of the Pleckstrin family of folds. Hubbard's fold predictions for Biphenyl-2,3-Diol 1,2-Dioxygenase and Synaptotagmin C2 were accomplished with a combination of secondary structure prediction and threading. For these reasons, it can be argued that these approaches to fold prediction should

be classified under the "Threading" category of structure prediction. *Ab-initio* fold prediction would then be limited to the strict combinatorial approaches to tertiary structure or methods that did not employ secondary structure units as intermediates. This may prove to be an arbitrary distinction. As *ab-initio* methods improve, threading algorithms will exploit these features to their own advantage. If there are only a limited number of protein folds, and X-ray crystallographers and NMR spectroscopists continue to solve a wide array of new structures, threading algorithms are likely to limit the need for true *ab-initio* approaches.

Primary → Tertiary

This category of methods includes those that do not use the calculation of secondary structure as an intermediate for structure determination. These include the work of Mekler, Marshall, Lee, Moult and Covell. These methods have the advantage over Primary → Secondary → Tertiary methods in that they are not simply threading. They are clearly applicable to the prediction of novel structures and folds. Moreover, the work of Chan and Dill(Chan & Dill, 1990) would suggest that secondary and tertiary structure are inextricably tied. This has led to the notion that tertiary structure determines secondary structures. While this extreme point of view is unlikely to be true in all cases (Gregoret & Cohen, 1991; Kim *et al.*, 1982), it is clear that the structure of some local sequences is largely influenced by their tertiary context (Cohen *et al.*, 1993; Kabsch & Sander, 1984).

Contact Matrices

Mekler and Marshall both used methods that involved generating the tertiary structure from a predicted set of interresidue contacts. Unfortunately,

as with the Primary → Secondary → Tertiary methods, Mekler and Marshall were unable to correctly predict the structure of Domain 3 of Staufen. Marshall incorrectly predicted the structure of Subtilisin propiece, and Mekler incorrectly predicted the fold of the Replication Terminator Protein. These results are not surprising in light of the predicted contact matrices of Mekler. For Domain 3 of Staufen, Mekler correctly predicted none of the seven long range (greater than five residue separation) contacts. For the Replication Terminator Protein, Mekler predicted none of the twenty-three long range contacts correctly.

Semi-Exhaustive Methods

The membrane-binding domain for the C2 domain of human coagulation factor VIII indicates that structures of very small proteins may be easier to predict. This peptide is only 22 amino acids in length. The NMR structure, and two lowest energy predicted structures are shown in **Figure IV.10**. The structures from both Moult's group and Lee's group are qualitatively quite accurate: a helix followed by an n-terminal twist. As shown in **Table IV.3**, each predicted structure deviated from the NMR structure by 4.4 Å r.m.s., and had low soap bubble values. Moult's group made two other predictions and Lee's group one other. These other predictions were higher in energy and correspondingly less accurate. However, Lee's high energy prediction was convincing enough that it was chosen by that group to be their preferred prediction.

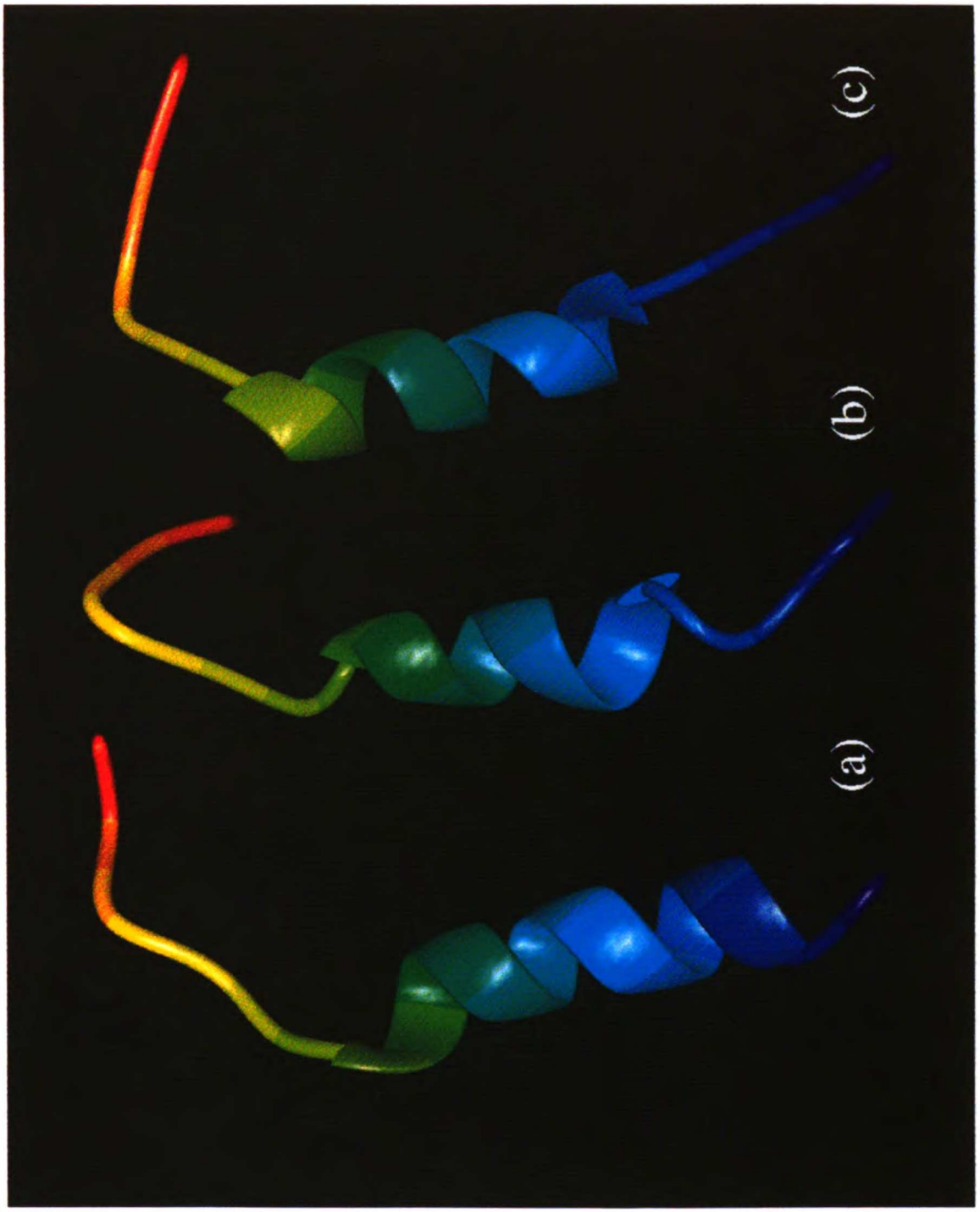


Figure IV.10. Membrane Binding domain for the C2 domain of human coagulation factor VIII. (*a*) NMR structure (*b*) Lee prediction and (*c*) Moult prediction. Pictures generated with MIDAS ribbonjr (Ferrin *et al.*, 1988) . The N-terminus is red, the C-terminus is blue, and the middle of the sequence is green.

One partially successful example of *ab-initio* protein folding was the effort of Covell on Chymotrypsin/Elastase Inhibitor-1. His predicted structure for this 65 residue protein was 7.3 Å r.m.s. from the actual structure. At the simplest level of comparison, this difference is ~3 standard deviations from an average random structure, but ~5 standard deviations from the actual structure (**Table IV.3**). The soap bubble value is indicative of a loosely similar structure. Visual inspection also shows some structural similarities between the experimental and predicted structure (**Figure IV.11**). Five disulfide bonds were present in this protein and the exact pairings were made known to the investigators. Covell did not use these pairing during the prediction phase of his work, but only later as a check on the accuracy of his prediction. Since he did have knowledge of these pairings before hand, this prediction cannot be considered entirely "blind". Still, disulfide bridge information is often available in advance of a structure determination and thus provides a useful type of experimental structure constraint for *ab-initio* methods (Cohen *et al.*, 1986b; Curtis *et al.*, 1991)

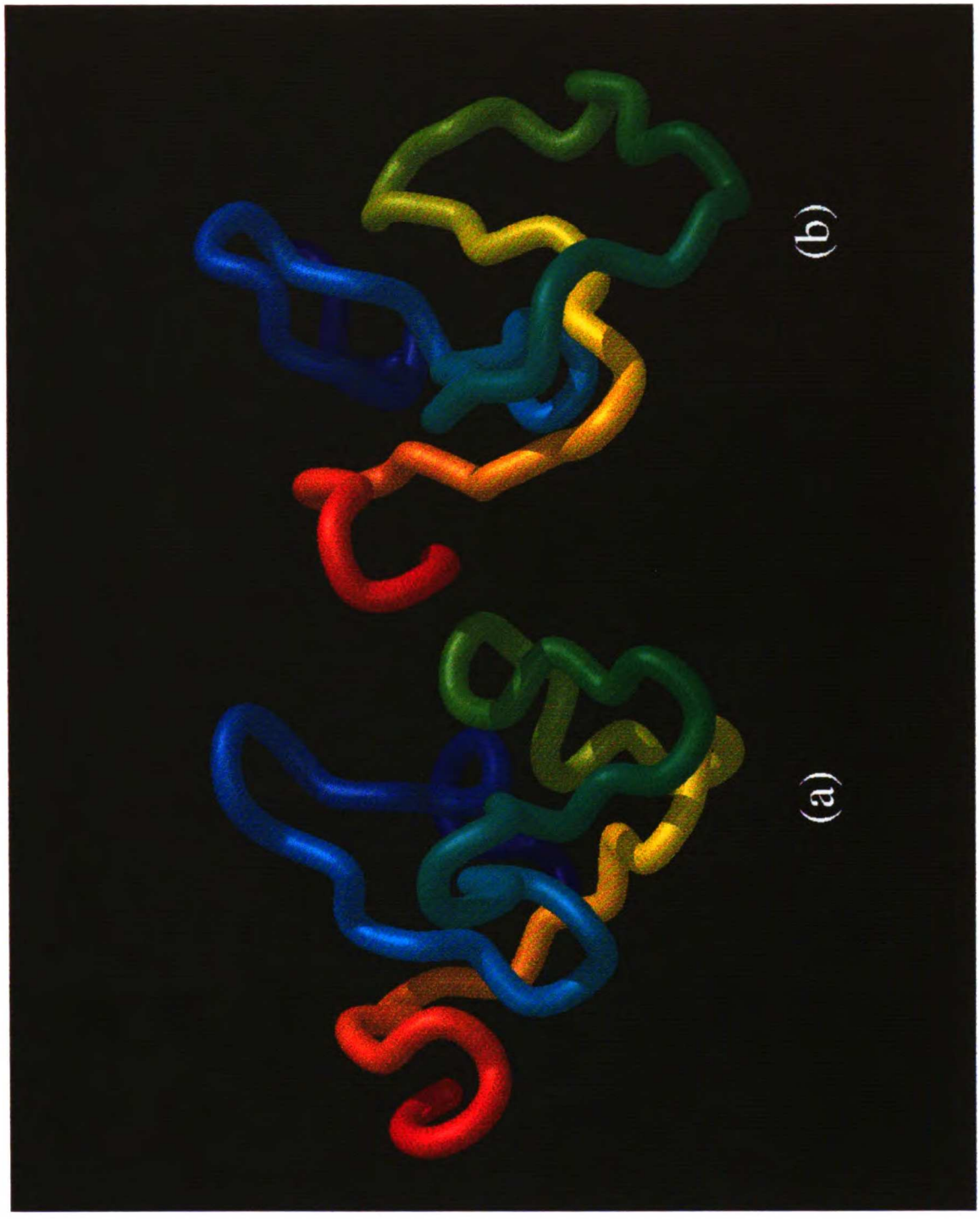


Figure IV.11. Chymotrypsin/Elastase Inhibitor (*a*) predicted and (*b*) experimental. Picture generated with Midas Neon (Ferrin *et al.*, 1988). α-carbon representation with exaggerated carbon radius used to emphasize the topological features. The tube was colored in a rainbow pattern corresponding to the amino acid number. The N-terminus is red, the C-terminus is blue, and the middle of the sequence is green.

Conclusion

Accurate tertiary structure prediction is not possible today. Overall fold prediction is possible and has been demonstrated. We can predict the overall fold of a protein when that protein has a recognizable motif. Examples are the leucine zipper seen in Chorismate Mutase and Replication Terminator Protein, and the α/β barrels Xylanase and 6-phospho-beta-D-galactosidase. We are also aided by a large degree of symmetry present in the amino acid sequence, which translates to symmetry in the three-dimensional structure. We can predict the approximate structure of extremely small proteins, such as the Membrane-binding domain for the C2 domain of human coagulation factor VIII. For these tiny proteins, it is possible to pursue extensive conformational searches to predict a protein's tertiary structure. Unfortunately, the predicted structures are still 4.4 Å r.m.s. from the experimental structures. There is hope that these methods may be extended to somewhat larger proteins as shown by Covell's prediction of Chymotrypsin/Elastase Inhibitor-1, but in this example, the structural resemblance is tenuous.

We still have difficulty with proteins that have unusual folding motifs, such as Domain 3 of Staufen and the Subtilisin Propiece. In addition, most of the recent advances made in structure predictions have been due to the exploitation of multiple sequence alignments. When the quality of these alignments was poor as was seen with the Replication Terminator Protein and Synaptotagmin I C2, prediction accuracy suffered. Given the current level of prediction accuracy, we recommend the use of as much experimental

information as possible in structure prediction and/or subsequent validation (Cohen *et al.*, 1986b; Cohen & Sternberg, 1980b; Jin *et al.*, 1994).

To improve tertiary structure prediction, multiple sequence information may have to be included in the Primary → Tertiary methods. Already, this information could improve the distance matrix approaches by using sequence variability information across an aligned family to narrow the range of coordination numbers for contact map approaches. Multiple sequence alignments could also be adapted to calculate more specific contact potentials. We expect that the next generation of tertiary structure prediction strategies will exploit multiple sequence information.

References

Avbelj, F. & Moult, J. (1995). Role of Electrostatic Screening in Determining Protein Main Chain Conformational Preferences. *Biochemistry* **34**, 755-764.

Avbelj, F. & Moult, J. (1995, in press). Determination of the Conformation of Folding Initiation Sites in Proteins by Computer Simulation. *Proteins*.

Benner, S. A., Badcoe, I., Cohen, M. A. & Gerloff, D. L. (1994). Bona Fide Prediction of Aspects of Protein Conformation. *J. Mol. Biol.* **235**, 926-958.

Benner, S. A., Cohen, M. A. & Gerloff, D. (1992). Correct structure prediction? *Nature* **359**, 781.

Benner, S. A. & Gerloff, D. (1991). Patterns of divergence in homologous proteins as indicators of secondary and tertiary structure: a prediction of the structure of the catalytic domain of protein kinases. *Advances in Enzyme Regulation* **31**, 121-181.

Benner, S. A. & Gerloff, D. (1993). Predicting the conformation of proteins. Man versus machine. *FEBS Letters* **325**, 29-33.

Biou, V., Gibrat, J. F., Levin, J. M., Robson, B. & Garnier, J. R. (1988). Secondary structure prediction: combination of three different methods. *Prot. Eng.* **2**, 185-191.

Bowie, J. U., Luthy, R. & Eisenberg, D. (1991). A Method to Identify Protein Sequences That Fold into a Known Three-Dimensional Structure. *Science* **253**, 164-169.

Bryant, S. H. & Lawrence, C. E. (1993). An Empirical Energy Function for Threading Protein Sequence Through the Folding Motif. *Proteins: Structure, Function and Genetics* **16**, 92-112.

Chan, H. S. & Dill, K. A. (1990). Origins of structure in globular proteins. *Proc. Natl. Acad. Sci., USA* **74**, 4130-4134.

Chothia, C. & Murzin, A. G. (1993). New folds for all-beta proteins. *Structure* **1**, 217-22.

Chou, P. Y. & Fasman, G. D. (1978). Prediction of the secondary structure of proteins from their amino acid sequence. *Advan. Enzymol.* **47**, 45-148.

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

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* **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.

Covell, D. G. (1992). Folding Protein alpha-carbon Chains into Compact Forms by Monte Carlo Methods. *Proteins* **14**, 192-204.

Covell, D. G. (1994). Low Resolution Models of Polypeptide Chain Collapse. *J. Mol. Biol.* **235**, 1032-43.

Crawford, I. P., Niermann, T. & Kirschner, K. (1987). Prediction of Secondary Structure by Evolutionary Comparison: Application to the alpha Subunit of Tryptophan Synthase. *Proteins* **2**, 118-129.

Curtis, B. M., Presnell, S. R., Srinivasan, S., Sassenfeld, H., Klinke, R., Jeffrey, E., Cosman, D., March, C. J. & Cohen, F. E. (1991). Experimental and theoretic studies of the 3-dimensional structure of human interleukin-4. *Proteins: Struct. Funct. Genet.* **11**, 111-119.

DiFrancesco, V., Munson, P. J. & Garnier, J. R. (1995). Use of Multiple Alignments in Protein Secondary Structure Prediction. *28th Hawaii*

International Conference on System Sciences, IEEE Computer Society Press, LosAlamitos **5**, 285-291.

Emery, S. C. & Fritz, H. J. (1994). Gene synthesis, expression and purification.

Falicov, A. & Cohen, F. E. (1996). Novel Metric for Structural Comparison of Proteins. *J. Mol. Biol.* **In Press**.

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

Galaktionov, S. G. & Marshall, G. R. (1994). Properties of Intraglobular Contacts in Proteins: An approach to Prediction of Tertiary Structure. *Proc. 27th Hawaiian International Conference on Systems Sciences, Biotechnology Computing, IEEE Computer Society Press* **5**, 326-335.

Garnier, J. R., Osguthorpe, D. J. & Robson, B. (1978). Analysis of the Accuracy and Implications of Simple Methods for Predicting the Secondary Structure of Globular Proteins. *J. Mol. Biol.* **120**, 97-120.

Godzik, A. & Skolnick, J. (1992). Sequence-structure matching in globular proteins: application to supersecondary and tertiary structure determination. *Proc. Nat. Acad. Sci. U.S.A.* **89**, 12098-102.

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

Gregoret, L. M. & Cohen, F. E. (1991). Effect of packing density on chain conformation. *J. Mol. Biol.* **219**, 109-122.

Harper, E. T. & Rose, G. D. (1993). Helix stop signals in proteins and peptides: the capping box. *Biochem.* **32**, 7605-9.

Harris, N. L., Presnell, S. R. & Cohen, F. E. (1994). Four Helix Bundle Diversity in Globular Proteins. *J. Mol. Biol.* **236**, 1356-68.

Hubbard, T. J. (1994). Use of β-strand Interaction Pseudo-Potentials in Protein Structure Prediction and Modelling. *Proceedings of the Biotechnology Computing Track, Protein Structure Prediction MiniTrack of the 27th HICSS. IEEE Computer Society Press*, 336-354.

Jin, L., Cohen, F. E. & Wells, J. A. (1994). Structure from function: screening structural models with functional data. *Proc. Natl. Acad. Sci. USA* **91**, 113-7.

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

Kabsch, W. & Sander, C. (1983). Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers* **22**, 2577-2637.

Kabsch, W. & Sander, C. (1984). On the Use of Sequence Homologies to Predict Protein Structure: Identical Pentapeptides can have Completely Different Conformations. *Proc. Natl. Acad. Sci., USA* **81**, 1075-1078.

Kang, H. S., Kurochkina, N. & Lee, B. K. (1993). Estimation and use of Protein Backbone Angle probabilities. *J. Mol. Biol.* **229**, 448-460.

Kim, P. S., Bierzynski, A. & Baldwin, R. L. (1982). A competing salt-bridge suppresses helix formation by the isolated C-peptide carboxylate of ribonuclease A. *J. Mol. Biol.* **162**, 187-99.

King, R. D. & Sternberg, M. J. (1990). Machine learning approach for the prediction of protein secondary structure. *J. Mol. Biol.* **216**, 441-457.

Kurochkina, N. & Lee, B. K. (1995, in press). Hydrophobic potential by pair-wise surface area sum. *Prot. Eng.*

Landschulz, W. H., Johnson, P. F. & McKnight, S. L. (1988). The leucine zipper: a hypothetical structure common to a new class of DNA binding proteins. *Science* **240**, 1759-64.

Leng, B. (1994). A Knowledge-Based Approach for Predicting the Internal Structure of Objects with Two-Level Case-Based Reasoning. PhD., U. of Pittsburgh.

Leng, B., Buchanan, B. G. & Nicholas, H. B. (1994). Protein Secondary Structure Prediction Using Two-Level Case-Based Reasoning. *Journal of Computational Biology* **1**, 25-38.

Levin, J. M. & Garnier, J. R. (1988). Improvements in a secondary structure prediction method based on a search for local sequence homologies and its use as a model building tool. *Biochim. Biophys. Acta.* **955**, 283-285.

Levin, J. M., Pascarella, S., Argos, P. & Garnier, J. R. (1993). Quantification of secondary structure prediction improvement using multiple alignments. *Prot. Eng.* **6**, 849-854.

Levitt, M. & Chothia, C. (1976). Structural patterns in globular proteins. *Nature* **261**, 552-558.

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

Mekler, L. B. & Idlis, R. G. (1993). The general stereochemical genetic code - the way to 21st-century biotechnology and universal medicine - already today. *Priroda (Nature) the monthly scientific journal of the Russian Academy of Science (in Russian; translation into English is available from the authors by request)* **5**, 22-25.

Munson, P. J., DiFrancesco, V. & Porrelli, R. (1994). Protein Secondary Structure Prediction using Periodic-Quadratic-Logistic Models: Theoretical and Practical Issues. *27th Annual Hawaii International Conference on System Sciences . IEEE Computer Society Press, LosAlamitos* **5**, 285-291.

Muskal, S. M. & Kim, S.-H. (1992). Predicting Protein Secondary Structure Content. *J. Mol. Biol.* **225**, 713-727.

Nishikawa, K., Kubota, Y. & Ooi, T. (1983). Classification of proteins into groups based on amino acid composition and other characters. I. Angular distribution. *J. Biochem.* **94**, 981-995.

O'Shea, E. K., Rutkowski, R. & Kim, P. S. (1989). Evidence that the leucine zipper is a coiled coil. *Science* **243**, 538-542.

Orengo, C. A., Flores, T. P., Taylor, W. R. & Thornton, J. M. (1993). Identification and classification of protein fold families. *Prot. Eng.* **6**, 485-500.

Pedersen, J. T. & Moult, J. (Document in preperation). Genetic Algorithms in Protein Folding: An efficient, full atom representation torsion space algorithm for the minimization of global energy functions. .

Ptitsyn, O. B. & Rashin, A. A. (1975). A model of myoglobin self-organization. *Biophys. Chem.* **3**, 1-20.

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

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

Ring, C. S. & Cohen, F. E. (1993). Modeling protein structures - construction and their applications. *FASEB J.* **7**, 783-790.

Robson, B. & Garnier, J. R. (1993). Protein structure prediction. *Nature* **361**, 506.

Rodionov, M. A. & Galaktionov, S. G. (1992). Analysis of the Three-Dimensional Structure of Proteins in Terms of Residue-Residue Contact Matrices. II. Coordination Numbers. *Mol. Biol.* **26**, 777-783.

Rost, B. & Sander, C. (1994). Combining Evolutionary Information and Neural Networks to Predict Protein Secondary Structure. *Proteins* **19**, 55-72.

Russell, R. B., Breed, J. & Barton, G. J. (1992). Conservation analysis and structure prediction of the SH2 family of phosphotyrosine binding domains. *FEBS Letters* **304**, 15-20.

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

Sander, C. & Vriend, G. (1992). .

Schmid, F. X. (1994). Spectra of RORO.

Schulz, G. E., Barry, C. D., Friedman, J., Chou, P. Y., Fasman, G. D., Finkelstein, A. V., Lim, V. I., Ptitsyn, O. B., Kabat, E. A., Wu, T. T., Levitt, M.,

Robson, B. & Nagano, K. (1974). Comparison of predicted and experimentally determined secondary structure of adenyl kinase. *Nature* **250**, 140-2.

Schulz, G. E. & Schirmer, R. H. (1979). Evaluation of Prediction Methods. In *Principles of Protein Structure*, pp. 122-8. Springer-Verlag, New York.

Sheridan, R. P., Dixon, J. S. & Venkataraghavan, R. (1985). Amino Acid Composition and Hydrophobicity Patterns of Protein Domains Correlate with Their Structures. *Biopolymers* **24**, 1995-2023.

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.

Summers, N. L. & Karplus, M. (1990). Modeling of globular proteins. A distance-based data search procedure for the construction of insertion/deletion regions and Pro----non-Pro mutations. *J. Mol. Biol.* **216**, 991-1016.

Zhou, H. X., Lyu, P., Wemmer, D. E. & Kallenbach, N. R. (1994). Alpha helix capping in synthetic model peptides by reciprocal side chain-main chain interactions: evidence for an N terminal "capping box". *Proteins* **18**, 1-7.

Chapter 5

Multiple Sequence Information for Threading Algorithms

This chapter has been submitted for publication in the Journal of Molecular Biology for 1996

Introduction

The solution to the inverse protein folding problem lies in identifying amino acid sequences that are compatible with a given protein structure. Beginning with the work of Eisenberg and coworkers (Bowie *et al.*, 1991), threading algorithms have emerged as a valuable tool for identifying structure-sequence correlations. While many threading algorithms have succeeded in associating protein folds with compatible sequences, a recent analysis of these methods suggests that the sequence alignments implicit in the "correct" threading of a sequence through a structure are frequently incorrect (Lemer *et al.*, 1995). Presumably, these errant alignments undermine the efficacy of existing threading algorithms.

Recent algorithmic advances in protein secondary structure prediction exploit the additional structural information inherent in a family of aligned amino acid sequences (Benner *et al.*, 1994; Rost & Sander, 1993). In large part, their success is due to the utility of these alignments in separating structurally relevant regions from incoherent sequence information by recognizing amino acid positions that are relatively conserved or more freely variable. In an effort to explore the utility of positional variability measures, we have developed a method with a simplified structural environment-sequence potential for threading (The Test of Optimal Mutagenesis or TOM) and compared this algorithm with a literature standard: THREADER (Jones *et al.*, 1992). This analysis demonstrated that the current standards for assessing the success of threading algorithms need to be made more rigorous.

Historically, threading methods have been evaluated using a modest set of test cases. A familiar example is to thread, without gaps, a protein

sequence on every structure equal to or longer than itself (Hendlich *et al.*, 1990). Another control involves swapping the sequence of a protein with that of a protein of equal length, and rating the scoring function's ability to distinguish the correct from the swapped structure (Novotny *et al.*, 1988) . Unfortunately, these test cases have a sufficiently large number of structure-sequence incompatibilities that simplistic threading algorithms identify the errant structures with perfect or near perfect predictability. To evaluate the relative performance of TOM versus THREADER, we developed a new test case derived from an exhaustive analysis of all of the structural matches within a unique set of protein structures. A summary of our approach is given in **Figure V.1**. We compared 113 sequences from large homologous families (≥ 15 members, where at least one member of each family is present in the PDB) with 305 high resolution structures. 56 of the sequences had at least one close structural match other than itself (or members of the same homologous family) in the set of 305 structures. Our definition of close structural matches is given in the "methods" section. The test case involves searching our database of structures with an amino acid sequence. The structure-sequence matches are ranked by score, and a correct "hit" for each test is recorded when the highest scoring non-identical structure-sequence pair has a corresponding structure-structure match. This test case mimics blind structure predictions akin to the Asilomar folding challenge (Lemer *et al.*, 1995). While this test set is not totally independent of the training set, we have used a jackknife data analysis strategy to minimize memorization effects.

This test case allowed us to differentiate the performance of TOM and THREADER. Due to the dissimilarities between these two algorithms, it is not possible to determine the extent to which the inclusion of multiple

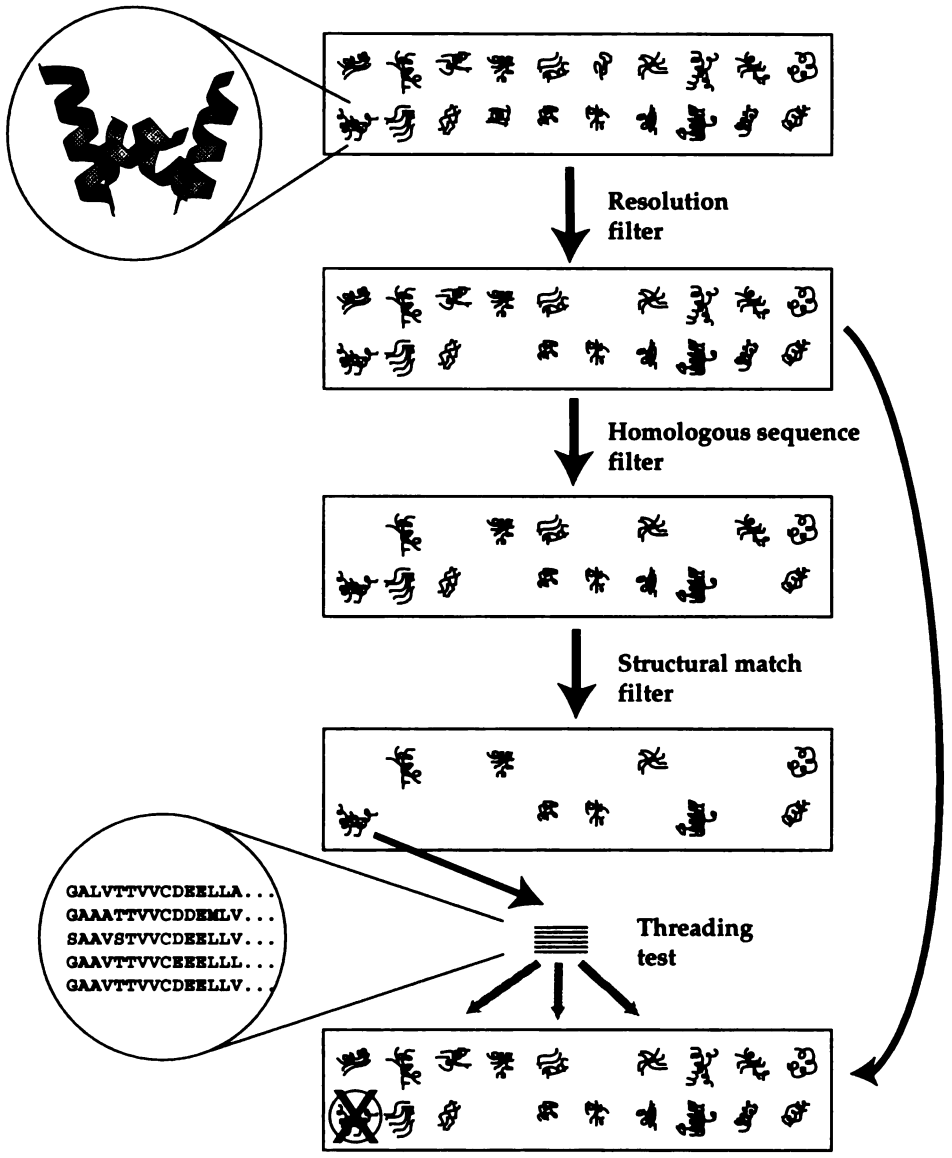


Figure V.1. A set of structures spanning the known protein folds (with less than 25% sequence identity to one another) is paired down to a test set of structures. The resolution filter eliminates all structures with less than 2.5 Å resolution. The resulting group represents our set of folds. With the homologous sequence filter this set is further paired down to include only those with at least 15 sequences in an associated multiple sequence alignment. The test set is generated with the structural match filter by keeping only those proteins that have structural matches other than themselves in the set of folds. The test case is to thread the sequence alignment from each protein in our test set against the set of folds. In each case, the highest scoring fold (excluding matching a sequence to its own native fold) is chosen. A more detailed description of each step is found in the "methods" section. The helical bundle was generated with MOLSCRIPT (Kraulis, 1991).

sequence information was responsible for the improvement. For this reason, an alternative version of TOM which does not employ variability information (TOM NOVAR), was developed in parallel with TOM. TOM NOVAR is otherwise identical to TOM. Contrasting TOM, TOM NOVAR and THREADER has revealed important insights into what makes threading algorithms succeed or fail.

Results and Discussion

The performance of TOM, TOM NOVAR, and THREADER on our test case is summarized in **Table V.1**. Clearly, multiple sequence information improved the performance of TOM relative to TOM NOVAR and THREADER. For each of the correct structure-sequence matches (or "hits") identified by TOM in our test case, we also evaluated the accuracy of the sequence alignments produced by both TOM and TOM NOVAR. TOM and TOM NOVAR were evaluated on the same set of "hits" in order to directly compare the algorithms. Unfortunately, since TOM NOVAR is being tested on the matches produced by TOM, it is possible that its performance is hampered. The alignments produced by both algorithms were compared to the alignments produced by an automated structure comparison algorithm (Falicov & Cohen, 1996). We counted the number of correctly aligned positions and divided by the number of residues aligned in the structural alignment. At the maximum stringency or "0" resolution level, the aligned positions had to be identical in the threaded and structural alignments. At the "±1" resolution level, a one residue discrepancy was still considered to be an acceptable alignment. Realistically, automated alignment algorithms that compare structures will differ at the "0" resolution level but will be concordant at the "±1" threshold. At the more modest "±3" or "±5"

Table V.1. Number of correct folds identified.

Algorithm	Fraction (%) of folds correctly matched
TOM	25/56 (45%)
TOM NOVAR	16/56 (29%)
THREADER	11/56 (20%)

resolution level, an acceptable alignment accommodated a disagreement of as many as 3 or 5 residues. The results for TOM and TOM NOVAR are summarized in **Table V.2**. The actual accuracy of TOM on a residue by residue percentage basis for each structure-sequence pair is shown in **Table V.3**. We believe that the stringency level of "±5" is approximately equivalent to the structural overlap method employed by Wodak and co-workers to review threading predictions presented at the Asilomar conference (Lemer *et al.*, 1995). At this level, the TOM algorithm predicted over 50% of the structural alignment correctly in 22 out of 25 cases. This compares favorably with the results presented at Asilomar; only Sippl and co-workers were more accurate on the limited set of three proteins that constituted this blinded challenge.

However, we believe that agreement at the "±1" resolution level more realistically represents the match between the sequence and the correct environmental positions in the fold. At this level of resolution, TOM correctly predicted 30% or more of the structural alignment for 21 out of 25 cases, while TOM NOVAR was successful in 16/25 cases. Multiple sequence information clearly improved the alignment accuracy. Interestingly, the number of cases that matched over more than 50% of the alignment was significantly decreased, just 15/25 and 13/25 for TOM and TOM NOVAR respectively. This leads us to believe that the structure-sequence matches identified by TOM require just 30-50% of the structure-sequence alignment to be correct in order to generate sufficient signal to distinguish it from alternate folds. These correctly aligned residues in some cases comprise one portion of the fold, and in other cases they are spread throughout the fold.

This work demonstratess an improvement over existing methods at two levels. First is the inclusion of additional information from mulitple

Table V.2. Summary of sequence alignment accuracy.

Stringency	Minimum % of sequence correctly aligned	Fraction (%) of alignments correctly predicted	
		TOM	TOM NOVAR
0	30%	14/25 (56%)	11/25(44%)
0	50%	6/25 (24%)	8/25 (32%)
0	90%	4/25 (16%)	4/25 (16%)
±1	30%	21/25 (84%)	16/25 (64%)
±1	50%	15/25 (60%)	13/25 (52%)
±1	90%	6/25 (24%)	5/25 (20%)
±3	30%	23/25 (92%)	20/25 (80%)
±3	50%	20/25 (80%)	18/25 (72%)
±3	90%	6/25 (24%)	6/25 (24%)
±5	30%	25/25 (100%)	22/25 (88%)
±5	50%	22/25 (88%)	20/25 (80%)
±5	90%	12/25 (48%)	11/25 (44%)

Table V.3. Sequence alignment percentage accuracy of TOM at three different stringency levels.

Sequence (PDB entry[1] + chain descriptor)	Structure (PDB entry + chain descriptor)	Stringency "0"	Stringency "±1"	Stringency "±3"	Stringency "±5"
1ads_	1ghsa	9	10	27	44
1apme	1irk_	35	45	49	50
1apve	1smra	32	69	84	100
1babb	2mge_	100	100	100	100
1cdta	3ebx_	0	18	55	80
1cpcb	1cpca	55	56	61	76
1eca_	3sdha	40	51	86	95
1gdm_	1babb	24	34	60	83
1hdca	1dhr_	31	31	63	86
1knt_	7pti_	98	98	98	98

[1]1aaj_ AMICYANIN; 1ads_ ALDOSE REDUCTASE; 1apme E chain of C-AMP DEPENDENT PROTEIN KINASE; 1apve E chain of ACID PROTEINASE; 1arb_ ACHROMOBACTER PROTEASE I; 1ars_ ASPARTATE AMINOTRANSFERASE; 1ash_ HEMOGLOBIN; 1babb B chain of HEMOGLOBIN; 1bet_ BETA-NERVE GROWTH FACTOR; 1bgeb B chain of GRANULOCYTE COLONY-STIMULATING FACTOR; 1bmda A chain of MALATE DEHYDROGENASE; 1caua A chain of CANAVALIN; 1caub B chain of CANAVALIN; 1ccr_ CYTOCHROME C; 1cde_ PHOSPHORIBOSYLGLYCINAMIDE FORMYLTRANSFERASE; 1cdta A chain of CARDIOTOXIN V4; 1cpca A chain of C-PHYCOCYANIN; 1cpcb B chain of C-PHYCOCYANIN; 1cus_ CUTINASE; 1dhr_ DIHYDROPTERIDINE REDUCTASE; 1dts_ DETHIOBIOTIN SYNTHASE; 1eca_ HEMOGLOBIN III; 1fbaa A chain of FRUCTOSE-1,6-BISPHOSPHATE ALDOLASE; 1fnc_ FERREDOXIN NADP+ OXIDOREDUCTASE; 1frpa A chain of FRUCTOSE-1,6-BISPHOSPHATASE; 1frub B chain of FC RECEPTOR (NEONATAL); 1gdna A chain of D-GLYCERATE DEHYDROGENASE; 1gdm_ LEGHEMOGLOBIN; 1ghsa A chain of 1,3-BETA-GLUCANASE (ISOZYME II); 1gky_ GUANYLATE KINASE; 1hdca A chain of 3-ALPHA, 20-BETA-HYDROXYSTEROID DEHYDROGENASE; 1hlb_ HEMOGLOBIN; 1hmv_ HHAI DNA METHYLTRANSFERAS; 1hnee E chain of ELASTASE; 1hucb B chain of CATHEPSIN B; 1huw_ HUMAN GROWTH HORMONE; 1ifc_ FATTY ACID BINDING PROTEIN; 1ilr1 1 chain of INTERLEUKIN 1 RECEPTOR ANTAGONIST PROTEIN; 1irk_ INSULIN RECEPTOR; 1ivd_ INFLUENZA A SUBTYPE N2 NEURAMINIDASE; 1knt_ COLLAGEN TYPE VI; 1ldm_ L-LACTATE DEHYDROGENASE; 1lki_ LEUKEMIA INHIBITORY FACTOR; 1lpe_ APOLIPOPROTEIN-E3; 1mina A chain of NITROGENASE MOLYBDENUM-IRON PROTEIN; 1minb B chain of NITROGENASE MOLYBDENUM-IRON PROTEIN; 1mup_ MAJOR URINARY PROTEIN; 1nar_ NARBONIN; 1ofv_ FLAVODOXIN; 1pfka A chain of PHOSPHOFRUCTOKINASE; 1php_ 3-PHOSPHOGLYCERATE KINASE; 1ppn_ PAPAIN; 1ptx_ SCORPION TOXIN II; 1rsy_ SYNAPTOTAGMIN I; 1smra A chain of RENIN; 1tgsi I chain of TRYPSINOGEN; 1tie_ TRYPSIN INHIBITOR; 1tml_ ENDO-1,4-BETA-D-GLUCANASE; 1tph1 1 chain of TRIOSEPHOSPHATE ISOMERASE; 1tpla A chain of TYROSINE PHENOL-LYASE; 1xnb_ XYLANASE; 1zaac C chain of ZIF268E; 2ak3a A chain of ADENYLATE KINASE ISOENZYME-3; 2alp_ ALPHA-LYTIC PROTEASE; 2avh_ 1,3-1,4-BETA-D-GLUCAN 4 GLUCANOHYDROLASE; 2azaa A chain of AZURIN; 2bbkh H chain of METHYLAMINE DEHYDROGENASE; 2ccva A chain of CYTOCHROME C'; 2cdv_ CYTOCHROME C3; 2cnd_ NADH-DEPENDENT NITRATE REDUCTASE; 2dri_ D-RIBOSE-BINDING PROTEIN; 2hbg_ HEMOGLOBIN; 2hhma A chain of INOSITOL MONOPHOSPHATASE; 2hmza A chain of HEMERYTHRIN; 2hpea A chain of HIV-2 PROTEASE; 2ihl_ PROTEIN G; 2mge_ MYOGLOBIN; 2mnr_ MANDELATE RACEMASE; 2mtac C chain of METHYLAMINE DEHYDROGENASE; 2pia_ PHTHALATE DIOXYGENASE REDUCTASE; 2rspb B chain of ROUS SARCOMA VIRUS PROTEASE; 2sil_ SIALIDASE; 2sn3_ NEUROTOXIN; 2tgi_ TRANSFORMING GROWTH FACTOR-BETA 2; 3aahb B chain of METHANOL DEHYDROGENASE; 3cd4_ CD4; 3chy_ CHE Y; 3dfr_ DIHYDROFOLATE REDUCTASE; 3ebx_ ERABUTOXIN; 3sdha A chain of HEMOGLOBIN I; 3sgbi I chain of PROTEINASE B; 4enl_ ENOLASE; 4fxn_ FLAVODOXIN; 5p21_ C-H-RAS P21 PROTEIN CATALYTIC DOMAIN; 6fabl L chain of ANTI-PHENYLARSONATE ANTIBODY; 7pti_ TRYPSIN INHIBITOR; 8atca A chain of ASPARTATE CARBAMOYLTRANSFERASE; 8fabb B chain of IMMUNOGLOBULIN IGG1 FAB FRAGMENT;

1ldm_	1bmda	26	62	83	90
1lpe_	2ccya	0	0	25	30
1mina	1minb	29	52	57	69
1ofv_	4fxn_	49	62	79	95
1ppn_	1hucb	33	41	52	71
1ptx_	2sn3_	25	71	85	100
1tgsi	3sgbi	2	92	94	94
2ak3a	1gky_	31	36	49	60
2cnd_	1fnc_	41	46	63	65
2mge_	1babb	94	99	100	100
2sn3_	1ptx_	56	85	85	100
3chy_	4fxn_	20	63	80	80
3sgbi	1tgsi	98	100	100	100
4enl_	1ptx_	3	18	33	38
7pti	1knt_	2	98	100	100

sequences which leads to more accurate threading. Second, the use of a carefully constructured test case is necessary to recognize this improved performance. A useful threading test set must be both fair and sufficiently difficult that not all algorithms can succeed on any individual case and no algorithm can succeed in all cases. Most extant test cases for threading algorithms fail to meet these criteria. For example, a common proof of relevance involves threading a sequence onto a group of structures equal to or longer than itself. The test is considered a success if the sequence recognizes its own structure. Without gaps, this is an example of a highly cooperative event for which almost every threading function in the literature can succeed. Also, the exact structure is likely to be compatible with its specific sequence from a variety of perspectives. For example, exact hydrogen bond potentials that emphasize the importance of side chain-side chain or side chain-main chain hydrogen bonds will quite easily find the exact structure for a sequence from a wide variety of alternatives. Unfortunately, these same potentials are not necessarily useful for the realistic threading problem where the correct structure is unknown. Finally, when gaps are allowed in a test case that contains an exact sequence structure match, it can be difficult to separate the utility of the threading algorithm from the importance of a stringent gap penalty; as the gap penalty is increased, the gapped threading problem converges to the ungapped threading problem, with an almost invariable increase in performance.

We sought a test set that was truly orthogonal to the training set. To approximate this goal, we used a non-redundant set of protein structures, each of which had 25% or less sequence identity to all other members of the set. In addition, we systematically removed the native structure from the training set before testing. In an effort to make the test set more challenging,

all structures even remotely structurally similar to the native structure were removed from the training set using the criteria of Falicov and Cohen (1996). This should help to avoid the possibility of memorizing aspects of the native structure implicit in the homologous structures present in the training set.

We used one set of gap parameters for the entire set of proteins in the test case. This appears to be an obvious step, but in practice, reparameterization for each protein within a test set is a common occurrence in the literature. Whenever a test of globin fold recognition and imunoglobulin fold recognition are carried out separately with different gap penalties, a limited reperamaterization has occurred. This can improve algorithm performance in a fashion that does not easily generalize to the unknown test cases. Similarly, a constant definition of a structural match in the form of an overlap score was used. A constant cutoff of structural similarity was chosen so that we could not influence our choice of a hit by subjective feelings of what should be a hit.

Finally, an abridged test case was generated with completely separate training and testing sets. This limits the statistical power of the test case, but improves the likelihood that the results are generalizable to novel sequences. 30 sequences were compared against 29 structures to identify the nearest non-native sequence-structure match. The results are summarized in **Table V.4**. Note that the scores are greatly improved relative to the standard test case (**Table V.1**). This is not due to a substantial increase in performance. The smaller number of alternative incorrect folds decreases the odds that an incorrect fold will receive a high score.

Structural Matches

An accurate, unbiased definition of what represents a structural match is required to develop a useful test set. Many methods have been developed

Table V.4. Number of correct folds identified for an abreviated test set

Algorithm	Fraction (%) of folds correctly matched
TOM	23/30 (77%)
TOM NOVAR	19/30 (63%)
THREADER	12/30 (40%)

to compare structures, beginning with the early computational work of Rossman and Argos (1976) and the taxonomic studies of Richardson (1981). Taylor and Orengo (1989) have used double dynamic programming approaches to structural alignments, but had to employ an empirical gap penalty. Falicov & Cohen (1996) developed a differential geometry algorithm that calculates the approximate surface of minimum area or "soap film" between two structures. This soap film is relatively insensitive to insertions and deletions (gaps), resulting in a method for comparing structures of different lengths. To avoid length artifacts, the surface area value is scaled for the length of the two proteins resulting in a ratio score. This structural comparison has demonstrated a high degree of concordance between what we and others believe to be structural matches (Falicov & Cohen, 1996), and is the algorithm we have chosen to use.

Two different criteria for structural matches were chosen. These were selected both by examining a dendrogram of structural similarity for concordance with previous taxonomic work, (**Figure V.2**) and individual structural comparisons for specific difficulties (**Figure V.3**). We determined that a ratio score of 0.2 was indicative of a close structural match, and 0.3 a weak but noticeable structural match. These ratio scores were applied uniformly to the protein structural comparisons that gave rise to the test case.

Algorithmic Issues

In order to analyze the effects of multiple sequence information on threading, we employed an otherwise straightforward algorithm similar to that used by Eisenberg and co-workers (Bowie *et al.*, 1991). This algorithm models the protein structure as a string of environments, and aligns the sequence with the structure using a dynamic programming matrix. The

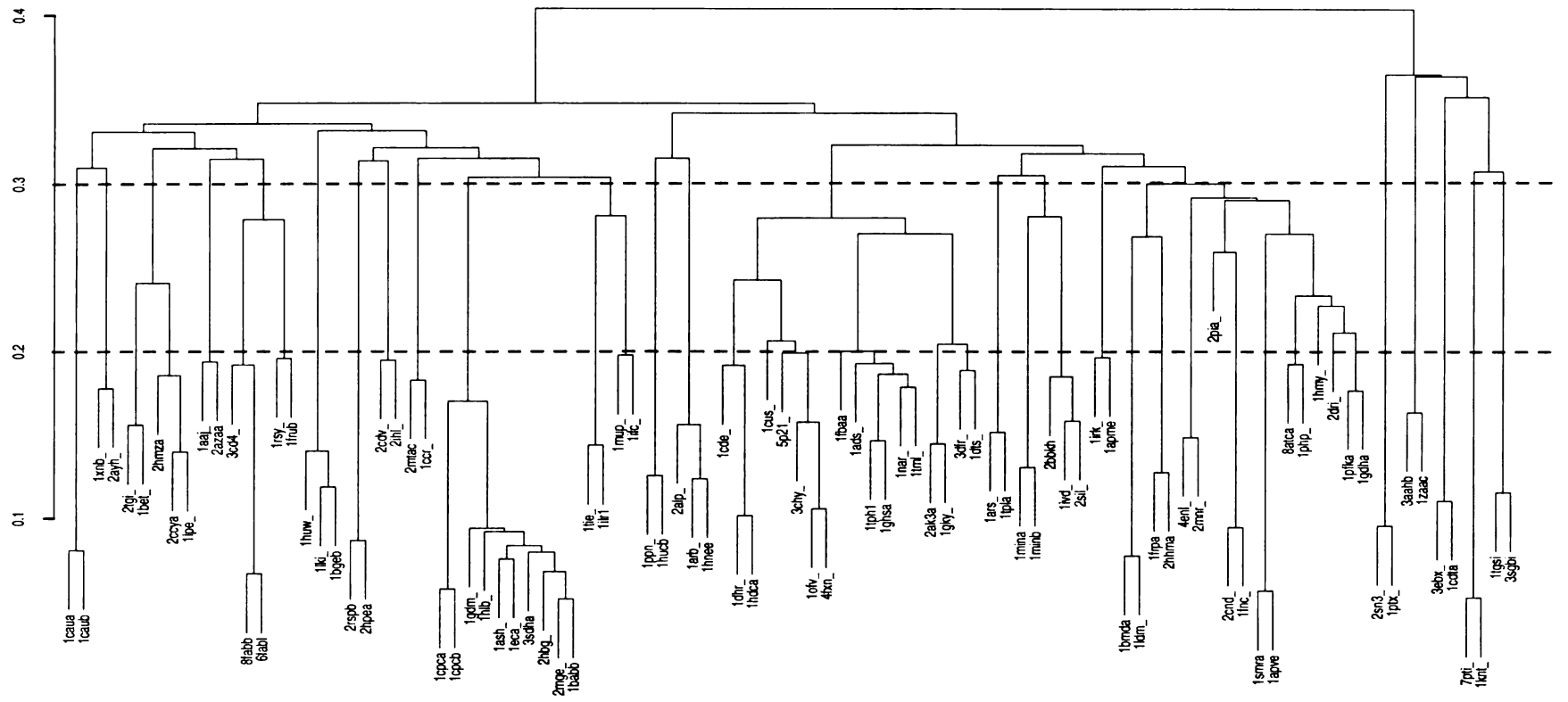


Figure V.2. Dendrogram of protein structure comparisons formed by calculating the structural similarity of each protein in a list to every other protein. The Soap Film method developed by Falicov and Cohen (1996) was used to make the comparisons. The list of proteins included the 56 proteins that formed the test set, and each of the structural matches to these 56, totaling 98 proteins. By analyzing a similar (much larger) dendrogram and individual comparisons such as in **Figure V.3**, we determined that a ratio score of 0.2 or lower was a close structural match, and 0.3 or lower a loose structural match. PDB entries are as specified in **Table V.3**.

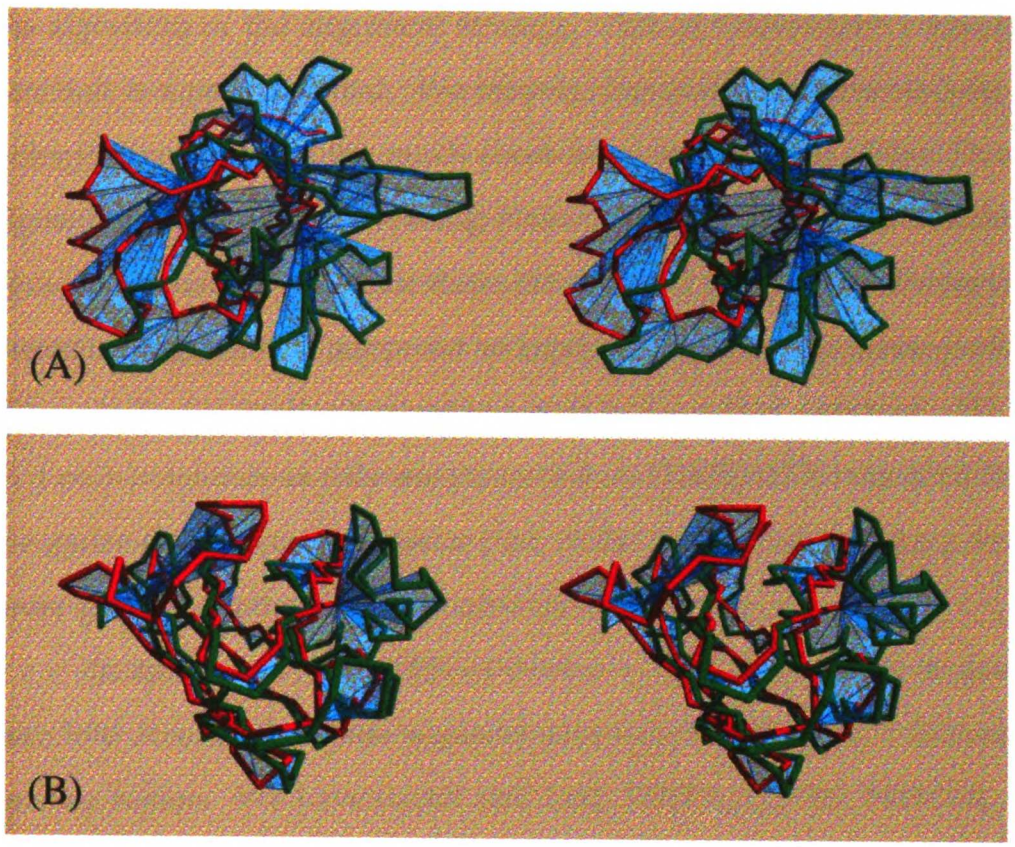


Figure V.3. Stereo view of two different protein-protein comparisons by the Soap Film method of Falicov and Cohen (1996). In (A), the two proteins are 1ilr 1 chain (green) and 1cew I chain (red). The ratio score is 0.288. In (B) the two proteins are 2aza A chain (green), and 1aaj (red). The ratio score is 0.194. We consider an ratio score of 0.2 or lower to be a close structural match, and 0.3 or lower to be an approximate structural match. PDB entries are as specified in the footnote to **Table V.3**.

structure-sequence matches are scored by assigning a value for each amino acid residue to be found in each environment.

The efficacy of this base algorithm is demonstrated by comparing TOM NOVAR and THREADER, as seen in **Table V.1.** On our threading test, TOM NOVAR outperformed THREADER by a 9% margin (29% to 20%). Since TOM NOVAR and THREADER are both based on matching a single sequence with a single structure, the difference in performance must be attributable to the algorithm or the alignment strategy. Part of THREADER's scoring function is a pair potential. We suspect that many implementations of pair-potentials incorporate too much short range information. Most threading algorithms derive their threading potentials by analyzing the matches between sequences and their three dimensional structures. However, a threading algorithm is supposed to locate a structure that is compatible with a sequence that is similar but not identical to the actual structure. We believe that pair-potentials as they are presented in THREADER may tend to recognize the structure-sequence pair, at the expense of capturing some of the more general similarities of an analogous structure.

In addition, THREADER calculates a "local" alignment. It does not take into account the entire sequence and structure; instead it focuses on a smaller region. We have found that both TOM and TOM NOVAR demonstrated substantial improvement with a "global" implementation of the dynamic programming algorithm. It is possible that this "global" alignment is responsible for much of the improvement seen in the structure-sequence alignments produced by TOM.

The Utility of Multiple Sequence Information

We took advantage of the information inherent in a family of aligned sequences by analyzing the variability of each position in the sequence.

Rather than focus on the overall variability of a sequence position, we subdivided the variability into hydrophobic and hydrophilic subsets. Benner and co-workers observed that extensive hydrophilic variability across the subset of hydrophilic amino acids is often associated with the surface exposure of a residue. In addition, they observed that a similar degree of variability restricted to hydrophobic residues with the explicit absence of hydrophilic residues is indicative of a buried side chain (Benner *et al.*, 1994). Thus, to perform the threading analysis we reclassified each amino acid in the sequence into one of four new amino acid types based on the hydrophobic and hydrophilic variability of the associated multiple sequence alignment (see methods section).

To calculate the variability of the sequence alignment, we classified each of the amino acids into one of three categories: hydrophobic, hydrophilic, and ambivalent. The hydrophobic amino acids all show a strong tendency to be buried. The hydrophilic amino acids show a strong tendency to be exposed. The ambivalent amino acids have approximately equivalent tendencies to be buried or exposed. Bowie *et al.* (1990) classify hydrophobics (Hφ) as {Phe, Ile, Leu, Met, Val, Trp, Cys}, hydrophilics (HP) as {Asp, Glu, Lys, Asn, Gln, Arg}, and ambivalents (HA) as {Ala, Cys, Gly, His, Pro, Ser, Thr, Tyr}.

We take issue with the classification of cysteine and serine. Cysteine behaves as a hydrophobic amino acid if it is part of a disulfide bond and behaves as a hydrophilic amino acid otherwise. In the absence of knowledge about the redox state of the protein, we believe that cysteine belongs in the ambivalent category. Serine also created a mild dilemma. From a study of multiple sequence alignments and protein structures, we concluded that the mutational rate, and the typical location of serine on the protein surface

exposed to solvent, placed serine in the hydrophilic category. Benner *et al.* (1994) also classified amino acids into three or four similar categories. They employed several different grouping in their papers, and some were quite similar to the ones used in the present work.

The evolutionary breadth of the multiple sequence alignment, and to some extent the actual number of sequences present in the associated multiple alignment, strongly influences the performance of TOM. During the variability calculation, the sequences are weighted so that alignments with different numbers of sequences produce the same overall signal. However, the precision of the variability calculated for a position within an alignment is a function of the number of sequences. When we tested families that had between 7 and 14 sequences in the associated multiple sequence alignment, we were able to locate a correct structure-sequence match in only 15% of the cases. By contrast, for sequences with more complete multiple sequence alignments, 45% of the correct matches were identified. In the absence of variability information (i.e. TOM NOVAR), 29% of the correct matches were identified. This result indicates that when variability information is far from complete, it may be more appropriate to ignore this information. When too few sequences are present in the alignment, or when those present are insufficiently different, the variability observed in a position may not be representative of the variability expected in a larger alignment. For example, in an alignment with 7 sequences, at a given position, variation amongst hydrophobics may be seen, indicating a tendency to be buried. However, a larger alignment might reveal a hydrophilic amino acid, changing this tendency towards exposure.

Conclusion

Central to our work is the value of an adequate test case. We believe that we were able to demonstrate the advantages of multiple sequence information due to the rigor of our test case. With this test case we will be able to evaluate future threading algorithms accurately, and thus gain insight into methods to improve the performance of threading algorithms.

By comparing TOM, TOM NOVAR and THREADER, we postulate that most current formulations of pair-potentials incorporate too much short range information. This information allows exact identification of the crystal structure, but can hamper the determination of similar folds. We have shown that algorithms based on the tendency of an amino acid to be buried or exposed (TOM NOVAR) are effective for threading. This is related to the conservation of patterns of burial and exposure across similar folds. The TOM algorithm demonstrates that the information in multiple sequence alignments can be used to improve the performance of threading algorithms.

Methods

Generation of multiple sequence alignments

The multiple sequence alignments used in the TOM method were generated by Sander and co-workers and were accessed from the HSSP database (Sander & Schneider, 1991). Each sequence is aligned to the sequence of a protein of known structure, and is predicted to fold to the same overall structure. New sequences added to these or other alignments are accepted or rejected based on whether their level of sequence similarity is greater than the threshold value t(L) in the following equation:

$$t(L) = 290.15L^{-0.562} + 3$$

where L is the length of the homologous region. This equation applies for L in the range of 10-80 residues. L values less than 10 are always rejected.

For L values greater than 80, t(L) is set at 28 percent (Sander & Schneider, 1991).

Calculation of variability

The information in the multiple sequence alignment is reduced to a one dimensional string of amino acids. In TOM, each standard amino acid in the multiple sequence alignment is broken down into one of four classes based on its hydrophobic and hydrophilic variability. Thus, the standard twenty amino acids are broken down into a total of 80 amino acid categories. The two hydrophobic classes are: Variable Hydrophobic $V_{H\phi}$ (Hydrophobic Variability >0.001) or Invariant Hydrophobic $I_{H\phi}$ (Hydrophobic Variability < 0.001). The two hydrophilic values are Variable Hydrophilic V_{HP} (Hydrophilic Variability > 0.001) or Invariant Hydrophilic I_{HP} (Hydrophilic Variability < 0.001).

Variability is calculated as:

$$\text{Hydrophobic Variability}(i) = \sum_{l=2}^{N} \sum_{k=1}^{l-1} \frac{\delta_{H\phi}(n_{ik}, n_{il}) * w_k * w_l}{d_{kl}}$$

$$\text{Hydrophilic Variability}(i) = \sum_{l=2}^{N} \sum_{k=1}^{l-1} \frac{\delta_{HP}(n_{ik}, n_{il}) * w_k * w_l}{d_{kl}}$$

$$\delta_{H\phi}(n_{ik}, n_{il}) = \begin{cases} 1 \text{ if } n_{ik} \neq n_{il} \text{ and } \begin{pmatrix} n_{ik} \in H\phi,\ n_{il} \in \overline{HP} \\ \text{or} \\ n_{ik} \in HA,\ n_{il} \in H\phi \end{pmatrix} \\ 0 \text{ otherwise} \end{cases}$$

$$\delta_{HP}(n_{ik}, n_{il}) = \begin{cases} 1 \text{ if } n_{ik} \neq n_{il} \text{ and } \begin{pmatrix} n_{ik} \in HP,\ n_{il} \in \overline{H\phi} \\ \text{or} \\ n_{ik} \in HA,\ n_{il} \in HP \end{pmatrix} \\ 0 \text{ otherwise} \end{cases}$$

Where N equals the number of sequences and n_{ik} is the ith amino acid of the kth sequence. d_{kl} is a measure of the evolutionary distance between the kth and lth sequences. We use the PAM distance, which is calculated from the sequence similarity of two sequences using a lookup table calculated by Dayhoff *et al.* (1972). w_k and w_l are the weights calculated for the kth and lth sequence. The sequences are weighted using the method described by Sibbald and Argos (1990). The weight of each sequence is approximately equal to the Voronoi volume of each sequence in sequence space, with the set of sequences weighted to unit weight. Hϕ is the set of hydrophobic amino acids {Phe, Ile, Leu, Met, Val, Trp}, HP is the set of hydrophilic amino acids {Asp, Glu, Lys, Asn, Gln, Arg, Ser}, HA is the set of ambivalent amino acids {Ala, Cys, Gly, His, Pro, Thr, Tyr}.

Protein Structure

The protein structure is reduced to a one dimensional string based on the percentage exposure of each residue (Exposed if ≥ 29%, Intermediate if < 29% and ≥ 7% and Buried if < 7%). The accessible surface areas were calculated using a modified version of the program ACCESS (Lee & Richards, 1971). The maximum exposure of each residue type was set to the maximum exposure value found from analyzing each protein structure from a non-redundant set of proteins published in a database by Hobohm *et al.* (1992).

Score Calculation

A table is calculated representing the compatibility of each amino acid type (generated from the original amino acid and the multiple sequence alignment), with each environment type (calculated from the structures using ACCESS). This table is generated by analyzing a non-redundant set of

proteins published in a database by Hobohm *et al.* (1992). The original list contains 453 proteins. We removed low resolution structures (>2.5 Å resolution), and structures with fewer than 15 sequences present in the associated multiple sequence alignment, resulting in the final list of 113 proteins. The score calculations are as follows:

score per table entry(i, j) =

$$\ln\left(\left(\frac{\#(\mathrm{aa}_i \text{ in } \mathrm{env}_j)}{\#env \sum_{l=1} \#(aa_i \text{ in } \mathrm{env}_l)}\right)\left(\frac{\#aa \sum_{k=1} \#(aa_k \text{ in } \mathrm{env}_j)}{\#aa \#env \sum_{k=1} \sum_{l=1} \#(aa_k \text{ in } \mathrm{env}_l)}\right)^{-1}\right)$$

Where aa_i is the ith amino acid in environment j (env_j).

This score is calculated from the sequences and environments of the proteins in the 113 protein list. This value represents how much more (or less) often an amino acid is present in an environment than a hypothetical "average" amino acid. A positive value represents a positive propensity for an environment. A separate table is calculated for each sequence in the test case. For each of these tables we remove the actual structure of the sequence, and any structure that is similar to the experimental structure, from the list of 113 sequences and structures.

Threading

Each sequence was compared to each structure using a simple dynamic programming matrix. This dynamic programming alignment allowed us, given constant gap and extension penalties, to find the optimal alignment between the string of amino acids representing the sequence and the string of environments representing the structure. The gap opening and extension penalties were empirically optimized to be 4.0 and 0.2 for TOM, 4.0 and 0.3 for

TOM NOVAR. The threading code is available by anonymous ftp at ftp.cmpharm in /pub/defay/threader.

Structural Comparison

The protein structure comparisons used to create the test sets were carried out with Soap Film (Falicov & Cohen, 1996). The cutoff values of 0.2 and 0.3 for structure similarity were determined by analyzing dendrograms similar to the one shown **Figure V.2**, and by observing individual structures such as those shown in **Figures V.3**.

Acknowledgments

This work was supported by the National Institutes of Health (GM 39900), and a grant from the National Science Foundation / Department of Energy (DE-AC03-76-SF01012)

References

Benner, S. A., Badcoe, I., Cohen, M. A. & Gerloff, D. L. (1994). Bona Fide Prediction of Aspects of Protein Conformation. *J. Mol. Biol.* **235**, 926-958.

Bowie, J. U., Clarke, N. D., Pabo, C. O. & Sauer, R. T. (1990). Identification of Protein Folds: Matching Hydrophobicity Patterns of Sequence Sets With Solvent Accessibility Patterns of Known Structures. *Proteins* **7**, 257-264.

Bowie, J. U., Luthy, R. & Eisenberg, D. (1991). A Method to Identify Protein Sequences That Fold into a Known Three-Dimensional Structure. *Science* **253**, 164-169.

Dayhoff, M. O., Hunt, L. T., McLaughlin, P. J. & Jones, D. D. (1972). *Gene duplications in evolution: the globins*, Silver Springs: National Biomedical Research Foundation.

Falicov, A. & Cohen, F. E. (1996). Novel Metric for Structural Comparison of Proteins. *J. Mol. Biol.* **Submitted**.

Hendlich, M., Lackner, P., Weitckus, S., Floeckner, H., Froschauer, R., Gottsbacher, K., Casari, G. & Sippl, M. J. (1990). Identification of native protein folds amongst a large number of incorrect models. The calculation of low energy conformations from potentials of mean force. *J. Mol. Biol.* **216**, 167-180.

Hobohm, U., Scharg, M., Schneider, R. & Sander, C. (1992). Selection of representative protein data sets. *Protein Science* **1**, 409-417.

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

Kraulis, P. (1991). MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Cryst.* **24**, 946-50.

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

Lemer, C. M. R., Rooman, M. J. & Wodak, S. J. (1995). Protein Structure Prediction By Threading Methods: Evaluation of Current Techniques. *Proteins* **23**, 337-355.

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

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

Rossman, M. G. & Argos, P. (1976). Exploring structural homology of proteins. *J. Mol. Biol.* **105**, 75-95.

Rost, B. & Sander, C. (1993). Prediction of Protein Secondary Structure at Better than 70% Accuracy. *J. Mol. Biol.* **232**, 584-599.

Sander, C. & Schneider, R. (1991). Database of Homology-Derived Protein Structures and the Structural Meaning of Sequence Alignment. *Proteins: Structure, Function and Genetics* **9**, 56-68.

Sibbald, P. R. & Argos, P. (1990). Weighting Aligned Protein or Nucleic Acid Sequences to Correct for Unequal Representation. *J. Mol. Biol.* **216**, 813-818.

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

The San Francisco Public Library

For reference

Not to be taken from the room.

627343

3 1378 00627 3430

San Francisco Public Library