# The Prediction and Analysis of Protein Structure Using Specialized Database Techniques

by

Tau-Mu Yi

Submitted to the Department of Biology
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 1995

Author ..........................................................................
Department of Biology
May 3, 1995

Certified by ....................................................................
Eric S. Lander
Professor of Biology
Thesis Supervisor

Accepted by ....................................................................
Frank Solomon
Chairman, Departmental Committee on Graduate Students

# The Prediction and Analysis of Protein Structure Using Specialized Database Techniques

by

Tau-Mu Yi

Submitted to the Department of Biology
on May 3, 1995, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

## Abstract

The subject of this thesis is the use of specialized database methods to analyze and predict protein structure. I have developed techniques for extracting and organizing the wealth of information in the protein structure and sequence databases. More specifically, the goal has been to identify and correlate sequence patterns with structural patterns. This approach has been applied to three general problems. First, the exhaustive pairwise structural comparison of known protein domains has led to the identification of shared structural topologies at the subdomain level (SSTs). I have provided evidence that these recurring substructures represent meaningful structural units and may form the basis of a new taxonomy for describing protein architecture. Secondly, I have used a nearest-neighbor classifier to predict secondary structure to an accuracy of 68% – among the best results to date. Moreover, I have devised a scheme to assign an *a priori* confidence level to each prediction, thereby distinguishing the more reliable predictions from the less reliable. The third project has been the development of an iterative procedure for predicting the fold of a protein. This method termed Iterative Template Refinement (ITR) involves the construction of templates combining structure-based and sequence-based information and employs an iterative search procedure to detect related proteins and to add them sequentially to the templates. On 6 test cases, ITR has demonstrated excellent sensitivity and selectivity.

Thesis Supervisor: Eric S. Lander
Title: Professor of Biology

# Acknowledgments

First, I would like to thank my thesis advisor Eric Lander for his support and guidance during the past five years. I learned many valuable lessons from my association with Eric, ranging from the use of mathematics to clarify one's analysis of a problem to the importance of communicating scientific findings to the general public. However, the most important aspect of my training has been learning how to complete a task in a timely and effective manner. There are few better able to teach project management skills than Eric.

Special thanks to Bruce Tidor for allowing me to attend his group meetings, journal clubs, afternoon teas, etc. I derived enormous benefit from the opportunity to interact with Bruce and his group. I learned a great deal about molecular mechanics, thermodynamics, and a variety of other issues in protein folding; but even more important was becoming integrated into a small community with similar interests as myself. It was not surprising that many people in the Whitehead thought that I belonged to the Tidor group. I will always remember Bruce's generosity to me.

I often told my friends that my thesis committee was the "Dream Team" of structural biology: Carl Pabo, Bob Sauer, Tomas Lozano-Perez and Peter Kim (thankfully, they charged much less than other Dream Teams). They have provided me with excellent advice over the years, and I hope to take advantage of their advice in the future.

I would also like to thank my first advisor Paul Schimmel for allowing me to pursue my interests even though they gradually led me away from his lab. I have a lot of respect for Paul and his sense of fairness and responsibility.

(Acknowlegments continued at end of thesis).

# Contents

# Chapter 1

# Introduction

In this introductory chapter, I will provide a general overview of protein structure, describe briefly the three problems investigated in this work, and outline the organization of the thesis. More specific background information will be found at the start of each chapter.

## 1.1 Protein Structure Primer

### 1.1.1 Basics

Proteins are important biological macromolecules that are responsible for catalyzing many of the chemical processes in the cell. Chemically, they are defined to be a polymer of amino acids linked together by peptide bonds. Proteins are comprised of 20 amino acids which differ in the structure of the side chain. A more thorough description of the properties of the amino acids and the peptide bond can be found in Creighton (1993). A typical polypeptide chain contains between 50 and 1000 residues or between 500 and 10,000 non-hydrogen atoms. The amino acid composition for most globular proteins does not deviate significantly from the average composition observed in the database, although there are notable exceptions (e.g., acid blobs, glutamine-rich activation domains, etc.). The 20 amino acid repertoire of proteins may be expanded by post-translational modification such as glycosylation, lipid attachment, sulfation, hydroxylation, and phosphorylation [Creighton, 1993].

Proteins adopt a unique three-dimensional conformation. In this respect, proteins differ from simple polymers in that they collapse to a single native structure instead of a large ensemble of compact conformations [Chan and Dill, 1991]. This structure is specified by the sequence of amino acids. Anfinsen demonstrated this point in a series of classic experiments in which he showed that the protein RNaseA could be denatured in urea and then refolded upon removal of the urea to its native conformation without the assistance of auxiliary factors [Anfinsen, 1973]. This simple view of protein folding has been complicated by recent findings that certain polypeptides require the presence of "molecular chaperones" to fold correctly [Hendrick and Hartl, 1993], but it is most likely that the chaperonins are kinetic catalysts of protein folding rather than active determinants of the final structure. Finally, although many proteins undergo conformational changes upon binding substrate or upon post-translational modification, these alterations typically involve subtle, local shifts in the structure ($<$ 1.0 Å for glycogen phosphorylase [Sprang et al., 1988]). Thus, each protein sequence is mapped onto one and only one protein structure.

The structure of a protein determines its function and regulation. Expediting the process of solving and analyzing new structures is a priority in molecular biology. Access to structural information is essential for scientists to achieve a detailed mechanistic understanding of biological systems. From the standpoint of drug design, an atomic resolution picture of the active site of a protein would aid pharmaceutical companies in their quest to design well-behaved inhibitors.

### 1.1.2 Experimental Methods for Protein Structure Determination

Experimental protein structure determination is both arduous and time-consuming. There are two main approaches for solving a protein structure: x-ray crystallography and NMR spectroscopy. The first step in x-ray crystallography is growing three-dimensional crystals of the purified protein. These crystals are placed into the path of an x-ray beam which scatters to form a diffraction pattern. Fourier transformation of this pattern in diffraction space results in an electron density map in Cartesian space. The phase of the reflections in the transform must be determined by either isomorphous replacement or molecular replacement. The final step is fitting a protein model into the electron density map. The quality of the model can be assessed by measuring the agreeement between the observed reflection amplitudes and the calculated amplitudes from the model (R factor). For more details on this important technique see the reference on x-ray crystallography by Stout and Jensen (1968).

Certain nuclei, most notably the nucleus of the hydrogen atom, possess a magnetic moment or spin that is oriented in an external magnetic field. NMR spectroscopy measures the resonance frequency of a radiofrequency pulse capable of flipping the spinning nuclei. These frequencies vary from nuclei to nuclei depending on the chemical environment, resulting in a one-dimensional pattern of resonance peaks along the frequency axis. In NMR studies of protein structure, the hydrogen nuclei in the molecule are typically monitored although two other less abundant isotypes that produce an NMR signal, $^{13}$C and $^{15}$N, can be artificially introduced into the protein. In order to obtain information about the structure of a protein, more sophisticated two-dimensional NMR spectra must be collected. In brief, it is possible to detect spin-spin coupling between two nuclei that are close in space ($\leq$ 5.0 Å). Measuring this interaction, termed the nuclear Overhauser effect (NOE), results in a set of distance constraints between different atoms in the molecule. One can then search for conformations of the protein that fit these constraints; usually, there are a small family of such structures (see Wutrich (1986) for more details). The principal advantage of NMR spectroscopy over x-ray crystallography is that the protein is studied in aqueous solution, a more natural setting, instead of being confined to a crystal lattice. Potential structural changes may be induced by crystal packing effects. On the other hand, x-ray diffraction can furnish a much more detailed picture of the protein.

Both techniques, however, presents some formidable technical obstacles ranging from growing crystals and obtaining heavy atom derivatives for x-ray crystallography to assigning resonance peaks and introducing NMR-active isotopes into the protein for NMR spectroscopy. These limitations explain why there are so few known protein structures (approximately 1000 non-identical structures [Orengo, 1994]). Moreover, some structures have proved refractory to solution by either method such as large complexes (e.g., the ribosome) and membrane proteins. Finally, although both crystallographic and NMR techniques have improved in recent years, a new structure still requires at least a few person-years to complete.

It is possible to obtain less complete structural information using lower resolution spectroscopic techniques such as circular dichroism (CD) and electron paramagnetic resonance (EPR) spectroscopy as well as chemical modification techniques such as cross-linking and immunochemical footprinting [Creighton, 1989, Sauer, 1993]. Ellipticity at 222 nm of a CD spectra provides a good measure of helical content in a protein. EPR probes can furnish information about the hydrophobic nature of the probe environment, the relative motion of the peptide segment attached to the probe, and the presence of neighboring probes. Cross-linking of functional groups belonging to different residues indicates close physical proximity, and treatment with monoclonal antibodies raised against the protein can map out regions of the polypeptide chain that are solvent exposed. These methods are well suited for monitoring the extent of protein folding under different experimental conditions or for studying the dynamic fluctuations in protein structure. On the other hand, they do not supply enough information to construct a high-resolution structure.

### 1.1.3 Theoretical Issues in Protein Folding

The forces, energetics, and thermodynamics of protein folding are incompletely understood. For small molecule model systems, chemists have been able to measure accurately thermodynamic parameters such as the enthalpy, entropy, and free energy for the formation of specific non-covalent interactions [Rigby et al., 1986]. Extrapolating these data to the larger and more complicated system represented by proteins is problematic. The principal hurdle is the the context dependence of small, higher-order effects that can be accurately modeled or simply ignored in a small compound, but must be approximated in a macromolecule. For example, a hydrogen bond between two water molecules *in vacuo* is known quite precisely to be worth −6.4 kcal/mol [Fersht, 1987], but the hydrogen bond between a serine and a glutamine residue is influenced by the environment of the bond. Electronic polarizability, reorientation of dipoles, and rearrangement of mobile ions are just some of the factors that could alter the strength of a hydrogen bond [Sharp and Honig, 1990]. An exact solution involving quantum mechanical calculations of the electronic distribution of the relevant atoms in the system is not feasible. Other unresolved questions concerning the forces that hold a protein together include the contribution of the hydrophobic effect to protein stability [Murphy et al., 1990], the interaction of helix macrodipoles with capping residues [Serrano and Fersht, 1989], and an estimation of the entropic loss for docking together two protein subunits [Tidor and Karplus, 1994].

Despite gaps in the theoretical understanding of proteins, substantial progress has been made in the field of computer simulations of polypeptide chains, i.e., molecular mechanics. Molecular mechanics (MM) attempts to simulate accurately the dynamics of a molecule using a chemically realistic energy potential (no quantum mechanical terms, however), subjecting the individual atoms to the various covalent and noncovalent forces, and solving Newton's equation of motions to determine the positions and velocities of the atoms [Karplus and McCammon, 1983]. The most important application of molecular mechanics to the study of protein folding has been the calculation of the free energy difference for a slight perturbation of the reference state – typically the native structure – to a new state by introducing a ligand, mutating a residue, etc. The values determined from these free energy calculations often agree quite closely to experimental numbers suggesting that the simulations capture the key interactions in the system [Lee, 1992]. The overall free energy difference can be broken down into components corresponding to specific parts of the protein or individual contributions of the energy function [Boresch et al., 1994]. This type of free energy analysis cannot be performed by an examination of the static structure.

On the other hand, molecular mechanics simulations are ill-suited for modelling the *de novo* folding of proteins. The difficulties arise from two sources: (1) finding the global free energy minimum structure is an extremely difficult optimization problem given the enormous size of the conformational search space and the limits on computational power. Put another way, a long MM simulation is on the order of 1 ns, whereas proteins fold on the time scale of 1 ms. The second basic problem is that the energy potential, although adequate for free energy perturbation calculations in which potential errors are largely cancelled, is not sufficiently accurate to distinguish the native conformation from the plethora of non-native structures [Novotny et al., 1984]. The absence of quantum mechanical terms, inaccuracies in energy parameters such as the dielectric constant, and an incomplete understanding of the physico-chemical interactions within a protein (see above) all contribute to an imperfect energy function. Thus, not only is it impossible currently to simulate the folding of a protein from an initial random coil state, but also a correctly folded protein will partially unfold in such simulations even under physiologic conditions [Daggett and Levitt, 1993].

Finally, one of the major unresolved theoretical questions in structural biology is the following: Does the native conformation of a protein correspond to the thermodynamic minimum free energy state (thermodynamic hypothesis) or are proteins caught in some metastable intermediate state (kinetic hypothesis)? In other words, when analyzing or predicting a protein structure, does one have to consider the kinetic pathway of protein folding. Dill has argued in support of the thermodynamic hypothesis based on experimental evidence that folding is thermodynamically reversible for a variety of single and multiple domain proteins suggesting that the native and unfolded states are in chemical equilibrium [Dill, 1990]. Moreover, lattice simulations by Shakhnovich and colleagues [Sali et al., 1994] have shown that although the protein does not have time to sample all possible conformations, a quick collapse to a condensed state followed by rearrangement of this collapsed globule to the thermodyanamically most stable state is kinetically reasonable. On the other hand, recent experiments with the influenza hemagglutinin protein has demonstrated that a shift to acid pH can induce a significant conformational change which is not reversed upon return to the starting conditions [Carr and Kim, 1993]. One possible explanation is that the native structure is in a metastable state and can only be released to a more stable final state by some environmental trigger. In summary, there is a large body of indirect evidence supporting the view that most proteins adopt the thermodynamically most favorable conformation, but there may be exceptions to this rule.

### 1.1.4 Hierarchy of Protein Structural Organization

The structure of a protein is quite complex. Proteins lack the regularity and symmetry found in DNA. Containing over 5000 atoms and 500 residues packed together in a compact globular arrangement, a typical protein defies a simple description. The task of analyzing and explaining protein structure has been facilitated by the recognition that protein structural organization is hierarchical. Structural biologists have distinguished seven levels of protein structure [Schulz and Schirmer, 1979, Jaenicke, 1991]: (1) primary structure, (2) secondary structure, (3) supersecondary structure, (4) subdomain structure, (5) domain structure, (6) tertiary structure, and (7) quaternary structure. At each level there are recurrent patterns or motifs which provide a basic vocabulary for describing the structural elements at that level. For example, at the secondary structure level, one can define regular local conformations of the backbone, alpha helices and beta strands, according to hydrogen-bonding patterns. Similarly, at the domain level, crystallographers have begun to assemble a taxonomy of structural folds that recur in a variety of proteins possessing different functions. Admittedly, the distinction between different structural levels may blur at times with some patterns seeming to belong to more than one level, but in general the hierarchy facilitates the parsing of a protein structure into more recognizable pieces.

Special display techniques have also assisted biologists in the interpretation and description of protein structure. Rather than attempting to visualize over 5000 spheres (atoms) and connecting line segments (bonds), one can choose to represent the structure in a more symbolic form using ribbon and topology diagrams [Flores et al., 1994]. This schematic depiction coupled with the hierarchical decomposition of the structure can help to unravel the myriad of atoms, bonds, and interactions in the protein.

### 1.1.5 Protein Sequence Information is a Valuable Resource

The difficulty of determining the structure of a protein, is matched only by the ease of obtaining the protein sequence. One can sequence a protein directly using a chemical technique called Edman degradation which involves the successive removal and identification of the amino terminal amino acid. The revolution in recombinant DNA technology has given rise to a much faster alternative strategy for protein sequencing. Namely, one can sequence the gene encoding the protein and then translate the DNA sequence into a protein sequence using the genetic code. Because of advances in automated DNA sequencing, it is now possible to sequence the cDNA encoding a protein of 500 amino acids in a single day. Currently, there are over 75,000 distinct entries in the OWL protein sequence database [Bleasby et al., 1994].

Significant sequence similarity is a strong indicator that two proteins are structurally and functionally related. Sequence similarity can be measured in terms of percent identity or a similarity score based on an amino acid substitution matrix. Dynamic programming is used to find the best alignment of two sequences given a scoring system and a set of gap penalties. Examining the Brookhaven structural databank, Lesk and Chothia have observed that proteins possessing greater than 20% sequence identity are very likely to be structurally homologous, and the degree of structural similarity varies directly with the degree of sequence similarity [Chothia and Lesk, 1986]. Indeed, one promising method for predicting the structure of an uncharacterized protein is to identify a related protein of known structure by sequence searching and then employing homology modelling to superimpose the new sequence on to the existing structure.

Because proteins related by sequence are likely to have descended from a common ancestor, they are also likely to possess similar catalytic activities. The first order of business after cloning a new gene of unknown function is to search the sequence database for homologs whose functional properties have been characterized. The immense size of the sequence database – over 50,000 entries and 20 million residues – and the relatively slow speed of dynamic programming have necessitated the development of fast heuristic methods for sequence comparison. The two most commonly used programs, FASTA [Pearson and Lipman, 1988] and BLAST [Altschul et al., 1990], can run a probe sequence against the database in a matter of minutes. The BLAST methodology has the further advantage of calculating the statistical significance of good matches. Recent data from the yeast genome sequencing project has revealed that approximately half of the newly sequenced genes register a significant hit against a sequence in the database [Koonin et al., 1994]. This 50% success rate has prompted research into more sensitive sequence searching techniques such as those using flexible pattern matching, multiple sequence templates, and libraries of motifs [Gribskov and Devereux, 1991].

The relative ease of obtaining the sequence of a protein and altering this sequence through *in vitro* or *in vivo* mutagenesis has opened a new avenue for probing the structure and function of proteins. Through site-directed mutagenesis, one can make single amino acid changes and evaluate their effect on stability or catalytic function. Alternatively, combinatorial mutagenesis schemes have enabled whole segments of the protein to be randomized. In this manner, one can exhaustively search for critical functional or structural determinants. One important conclusion from these studies is that proteins are much more tolerant to mutations than previously expected [Lim and Sauer, 1989]. Only a very few residues are absolutely essential for maintaining the native structure. This surprising plasticity presents a severe challenge to any attempt to describe the packing of a protein interior in terms of a set of rules. A rule-based approach fails because it cannot capture the range and diversity of interactions in proteins.

Finally, although proteins possessing similar sequences (i.e., $\geq 25\%$ sequence identity) invariably resemble one another in structure, the converse is not always true. Indeed, it has been one of the surprises of structural biology that proteins with very different sequences and functional properties could share the same basic structural topology (e.g., the RMS of TIM barrel structures with dissimilar sequences may be as low as 2.5 Å over greater than 100 positions). Thus, proteins that belong to the same fold family in structure space may be distantly related in sequence space.

## 1.2 Background to Thesis Topics

In this thesis, I have investigated three topics relating to the prediction and analysis of protein structure. The topics, covered in the middle three chapters of the thesis, are directed toward different levels of the protein structure hierarchy. Chapter 2 is concerned with the analysis of structural patterns at the subdomain level. Chapter 3 discusses the implementation of a nearest-neighbor classifier to predict secondary structure. Chapter 4 describes an iterative method for identifying sequences likely to adopt a specific fold (inverted protein structure prediction). Below, I provide general background information on each of the subjects.

### 1.2.1 Taxonomy of Protein Topologies

The domain constitutes a fundamental unit of protein structure. The concept of the domain originated when protein chemists noticed that proteolytic treatment of certain proteins resulted in the division of the functional activities of the protein amongst a set of stable fragments. Sequencing of these fragments revealed that they corresponded to compact, self-contained structural modules in the overall structure [Kirschner and Bisswanger, 1976]. Indeed, criteria have been established for dissecting a protein structure into component domains based on the ratio of solvent-exposed surface area to the surface area in contact with other parts of the protein [Holm and Sander, 1994]. The size of the average domain is between 60 and 500 residues; many proteins contain two or more domains.

In recent years, the structural biology community has been repeatedly surprised by reports that some new protein structure contains a domain that is structurally similiar to several other domains in the Brookhaven data bank despite the absence of significant sequence similarity. This finding that there are recurrent structural patterns at the domain level has given rise to speculation that there may a relatively small number of distinct domain topologies or folds [Chothia, 1993]. Some folds seem to be derived from a common ancestor (divergent evolution), but the passage of time has removed any trace of sequence or functional relationship. Other folds seem to be examples of convergent evolution in which the descendents of different ancestral proteins have converged on some favored packing arrangement [Lesk et al., 1989]. In either case, structural biologists have engaged in the task of grouping related domain structures into fold families and superfamilies. Jane Richardson initiated this endeavor by noting the patterns in $\beta$-strand connectivity in $\beta$-sheet proteins [Richardson, 1981]. Her work has been extended by a more systematic comparison of all structures in the database against one another and the clustering of similar structures into families [Orengo et al., 1993].

Completing the all-against-all comparison of database structures necessitated the development of fast and sensitive structural alignment programs. Formerly, calculating the optimal alignment and root-mean-square deviation of superimposed atoms (RMSD) between two structures required several minutes; the new generation of programs accomplishes this task in a matter of seconds. The key innovation has been the implementation of more sophisticated alignment algorithms ranging from double dynamic programming [Taylor and Orengo, 1989] to Monte Carlo optimization [Holm and Sander, 1993] to specialized graph searching [Mitchell et al., 1989] techniques. Another important technical advance has been the introduction of more robust measures of structural similarity. Root-mean-square deviation of superimposed atoms (RMSD) and alpha carbon distances (DRMS), the two most frequently used statistics for structural relatedness, are not properly normalized for different alignment lengths and are quite sensitive to shifts in the relative positioning of parts of the structures (e.g., the hinge motion of two domains). The SSAP statistic developed by Orengo and Taylor [Orengo et al., 1992] and 'elastic similarity score' of Holm and Sander [Holm and Sander, 1993] attempt to remedy these shortcomings.

How many fold families are there? The estimates range from 500 [Blundell and Johnson, 1993] to 8000 [Orengo, 1994]. The reason for this wide spread is that only a relatively small number of structures have been solved and this modest set suffers from sample bias (e.g., a disproportionate number of DNA-binding proteins and glycolytic enzymes have been crystallized). Extrapolating from this limited data is problematic. Moreover, it is now clear that the fold families are not equally populated. Orengo has shown that the 9 biggest 'superfolds' (e.g., TIM barrel, immunoglublin, etc.) account for 32% of the structures in the database, whereas the remainder of the database is distributed among 71 fold families [Orengo, 1994]. Finally, the very definition of a fold family suffers from some fuzziness. An arbitrary cutoff determines whether or not a domain belongs to a particular family, and sometimes a structure will fall just below or just above this threshold.

Yee and Dill have delved deeper into this issue by questioning the very concept of the fold family [Yee and Dill, 1993]. They studied whether the structural relationship between members of the same family was substantially closer than the relationship between members of different families. If fold families are tightly-knit entities, one would expect to see a bi-modal distribution of similarity scores in which one peak (small) corresponds to intra-family comparisons and one peak (large) corresponds to inter-family comparisons. Instead, a single continuous peak was observed, leading to the assertion that fold families are loosely-knit organizational units. This surprising conclusion, however, was potentially undermined by inadequacies in their structural comparison routine, i.e., no gaps or insertions were permitted.

The appearance of cracks in the 'fold-centric' view of the structural world has coincided with a renewed interest by structural biologists in the subdomain level of protein structure. Subdomains are most simply defined as compact pieces of domains that form distinct structural entities. They fall between supersecondary structure elements and domains in the hierarchy of protein structure. Because domains can be as large as 400 or 500 residues, experimentalists have attempted to identify subregions of the domain that can form autonomous folding units. In certain cases, such stable subdomains have been isolated [Vita et al., 1989, Jaenicke, 1991]. In addition, work characterizing the protein folding pathway has detected specfic subdomains as early folding intermediates [Hughson et al., 1990]. A picture of protein structure complementing the domain-level description is emerging from these studies of subdomains.

### 1.2.2 Secondary Structure Prediction

On the basis of model building and x-ray diffraction data on short peptides, Pauling et al. (1951) deduced that the polypeptide backbone adopts certain regular arrangements: $\alpha$-helices and $\beta$-sheets. These regular local conformations or secondary structures are stabilized by hydrogen bonds between the amide nitrogen and the carbonyl oxygen atoms. The $\alpha$-helix has a periodicity of 3.6 residues (right-handed) and the hydrogen bonds are between residues within a single helix. The $\beta$-sheet has a periodicity of 2.0 residues and the hydrogen bonds are between residues in two different elements or strands. Both parallel and anti-parallel hydrogen bonding patterns are observed in $\beta$-sheets. The third, default category of secondary structure, coil, includes the large number of irregular local conformations as well as the rare $3_{10}$- and $\pi$-helices. Approximately, 25% of residues are part of $\alpha$-helices, 20% are members of $\beta$-sheets, and the remaining 55% are "coil" positions.

The DSSP program written by Kabsch and Sander (1983a) has become the standard for assigning secondary structure given a coordinate file. DSSP uses hydrogen bonding patterns to determine the secondary structure type at a given position. Alternative methods use distance constraints between alpha carbon atoms (e.g., $i$ and $(i+4)$) [Levitt and Greer, 1977] or structural similiarity to canonical secondary structure elements to determine secondary structure. These different programs produce roughly equivalent outputs ($\geq$ 90% agreement).

Predicting secondary structure from sequence qualifies as a prototypical classification problem. Given a set of features, the sequence of the protein, one attempts to determine the labels (secondary structures) associated with the various positions in the protein. More specifically, the classifier attempts to determine the secondary structure label at position $i$ by examining a window of $n$ residues (typically between 9 and 21) centered at $i$. A range of techniques have been developed from the statistics and artificial intelligence (AI) communities for such problems. Many of these approaches have been applied to secondary structure prediction, including rule-based methods [Chou and Fasman, 1974], pattern-matching [Cohen et al., 1983], neural networks [Qian and Sejnowski, 1988], statistical methods [Gibrat et al., 1987], and database-oriented techniques [Levin et al., 1986].

To achieve peak performance, the internal parameters of the classifier must be adjusted or 'trained' over some part of the data. Thus, the data set is divided into multiple, non-overlapping test and training sets. The union of the test sets is the whole data set. This testing procedure, termed cross-validation, ensures that one does not test and train over the same data and that each member of the data set is tested exactly once. Only recently have researchers abided by this protocol, casting some doubt on previous claims [Kabsch and Sander, 1983b]. A second important check is that there should not be strong sequence homology between polypetide chains in the training set and chains in the test set. Zhang et al. (1992) removed any polypeptide chains in the database that possessed greater than 50% sequence identity with another member; Rost and Sander used the more conservative cutoff of 25% identity [Rost and Sander, 1993]. Again, this precaution has not been carefully observed by researchers evaluating secondary structure predictors.

The most commonly used statistic for prediction performance has been the percentage of positions correctly predicted. One drawback of this measure is that a 'dumb' predictor that predicts all coil positions achieves a prediction performance of 55%. As comparison, most current prediction systems successfully predict between 60 - 70% of residues. Matthews (1975) devised a measure based on the correlation between the predictions and the actual secondary structure that penalizes both false positives and false negatives so that an all-coil prediction would have a correlation score of 0. This statistic is particularly useful for assessing the effectiveness of the predictor over each of the secondary structure types. Finally, Rost and Sander (1993) and Yi and Lander (1993) have formulated a performance metric based on the amount of information provided by the predictions about the true secondary structure. The benefits of the information measure include its more precise mathematical foundation and its more intuitive representation of accuracy in terms of bits per position.

Rost and Sander have added an exciting new wrinkle to the enterprise of secondary structure prediction [Rost and Sander, 1993]. They have proposed to predict over multiple, aligned sequences instead of a single sequence. Because of the extra sequences, loop positions are more readily identified by the presence of gaps in the alignment, and the secondary structure 'signal' is more clearly distinguished from the background noise. The authors employed a cascaded neural network scheme in which the input encoding had been altered so that the proportion of each amino acid at a position was recorded instead of its presence or absence. The improvement was substantial; the neural network achieved a prediction accuracy of 71%. Not surprisingly, there was a direct correlation between the number of aligned sequences and performance.

Some have questioned the usefulness of secondary structure prediction. Knowing the secondary structure of a protein does not specify its complete 3D structure. I offer the following three motivations for continuing research in this area. First, information from these predictions can be used to make higher-order predictions by being incorporated into lattice model simulations or fold recognition programs. Secondly, the predictions provide insight into the factors that induce secondary structure formation (e.g., intrinsic amino acid propensities versus global packing effects). Finally, secondary structure prediction is a well-defined, yet challenging problem that serves as a testing ground for alternative prediction strategies.

### 1.2.3 Inverted Protein Structure Prediction

The possibility that there are a relatively small number of distinct protein folds suggests an indirect approach to the protein folding problem: assemble a database of folds and develop a classifier capable of assigning a protein sequence into the appropriate fold category and correctly aligning the sequence onto the fold. Alternatively, one could start with a given structural fold and search the database for sequences that are likely to adopt that structure. The two approaches are essentially equivalent and involve assessing the compatibility of a sequence with a structure. This inverted strategy of predicting structure has been referred to variously as 'inverse protein folding', inverted protein structure prediction, fold recognition, or threading. For a higher resolution model, one can take the canonical fold with the superimposed sequence and perform molecular mechanics to refine the hybrid structure.

Roughly speaking, one can divide the different approaches to inverted protein structure prediction into sequence-based methods and structure-based methods. Much of the current excitement surrounds the structure-based strategy, but for years scientists have employed sequence searching techniques to identify structurally related proteins. In its simplest form, the search template is composed of a single query sequence that is run against the database using dynamic programming or one of the fast heuristic comparison techniques such as FASTA and BLAST. The statistical significance of a match is assessed by permutation (shuffle) tests or by analytic equations derived from assumptions about the distribution of scores [Karlin et al., 1991]. To address the complaint that these single sequence comparison techniques lack sensitivity, researchers have developed more sophisticated techniques for exploiting sequence information that employ multiple sequence templates [Taylor, 1986], flexible pattern matching [Barton and Sternberg, 1990], and motif searching [Neuwald and Green, 1994].

The structure-based methods can be further divided into two camps: those that incorporate structural information into a local environment template and those that encode structural information into a residue-residue contact potential. Bowie and Eisenberg (1991) pioneered the former strategy by defining a set of 18 local structural environment classes based on secondary structure, solvent accessibility, and polarity. A scoring table was then established that assigned a value for pairing each of the amino acids with each of the environment classes. Dynamic programming found the optimal alignment of a target sequence onto the local environment template. Jones et al. (1992) were among the first groups to adopt the latter approach of transforming the structural data into a contact potential that measured the propensity for two residues to be in close proximity. The motivation is that the essence of a 3D structure can be captured by the set of contacts between neighboring residues, and the compatibility of a sequence with a structure can be estimated by computing the contact energy of the sequence aligned onto the structure. Because the contact score of a given residue depends on the alignment at the other positions, a variation of dynamic programming, double dynamic programming, was used to identify the best alignment. The many threading techniques that have appeared subsequently use variations of the local environment scoring system [Ouzounis et al., 1993], the pairwise contact potential [Sippl and Weitckus, 1992, Maiorov and Crippen, 1992], or both [Godzik and Skolnick, 1992, Goldstein et al., 1992]. Only Pickett et al. (1992) have merged structural and sequence information into a single template.

Although conceptually similar, the structure-based methodologies differ from one another in numerous technical details. Some of the more important technical issues include the following: (1) the scoring system for matching a particular structural feature with a particular amino acid; (2) the treatment of gaps and insertions; (3) computing the significance of a score; (4) weighting features from a single or multiple structures; and (5) the alignment method. Many of these implementation decisions have a profound effect on the overall performance of the program, and a more thorough understanding of their effects is needed.

The fold prediction algorithms have achieved some notable results. There is a long list of structurally related protein pairs possessing little or no sequence similarity that have been recognized by these techniques. Members of this list include actin and 70-kD heat-shock cognate protein [Bowie et al., 1991], globins and phycocyanin [Jones et al., 1992], hexokinase and actin [Jones et al., 1992], and plastocyanin and the immunoglobulins [Godzik and Skolnick, 1992]. While encouraging, these results have not been extended to a more comprehensive collection of test cases. Indeed, very few groups (Jones et al. (1992) is an exception) have compiled statistics – percent accuracy, number of false negatives, number of false positives, etc. – on the performance of their fold predictor. Thus, unlike the secondary structure prediction field, it is not possible to compare two methods based on some performance statistic. Developing more rigorous training and testing criteria is a priority.

As more protein structures are determined and the universe of structural topologies becomes more densely populated, inverted protein structure prediction will become an increasingly effective tool for predicting structure and function. Indeed, one possible 'solution' to the protein folding problem is the eventual identification of all distinct protein folds followed by the construction of a comprehensive database of fold templates, and the design of an accurate fold recognition program. This goal can be attained in the not-too-distant future.

## 1.3 Overview of Thesis

Managing the complexity inherent in protein structure is a key problem facing structural biologists. One seeks to organize and integrate the disparate collection of structural information. I have applied specialized database methods to address this problem. The basic idea is to use existing knowledge as examples to interpret new instances. By database techniques, I am referring to a broad range of methods, not only for storing and retrieving data, but also for processing and making sense of the data, i.e., comparing entries, finding patterns, quantifying the information in the database, etc. By manipulating data and patterns at different levels of complexity, this approach avoids the pitfall of being either too detailed (molecular mechanics) or not detailed enough (rule-based methods).

In this work, I have concentrated my effort on exploiting the vast amount of information in the sequence and structure databases. My basic strategy has been to identify sequence and structural patterns, correlate the two, and quantify the information present in these patterns. I define a pattern to be a concise description of a set of related data. Scanning the two databases for such patterns requires among other things a measure for comparing distinct entries, a method for clustering related entries, and a scheme for representing the combined features in a cluster. Moreover, it is possible to take advantage of the correlations between sequence and structural patterns to make predictions. For example, the sequence motif GxxxxGK[ST] is frequently associated with the beta-turn-alpha structural motif in GTP-binding proteins and can be used to predict the presence of a GTP-binding loop. Finally, I have employed some basic concepts from information theory to build a framework for assessing the effectiveness and reliability of both the patterns and the predictions.

I have explored three different topics from structural biology, described in the previous section, using this database approach. The three topics represent chapters 2, 3, and 4 of the thesis. In the final chapter, chapter 5, I tie together the results from this work.

Chapter 2 describes a project to identify structural patterns at the subdomain level. I have performed an exhaustive comparison of a database of structural domains taken from the Brookhaven databank. The important technical innovations include a length-normalized measure of structural similarity and a relaxed minimum alignment length criterion. The common substructures between two domains are termed shared structural topologies or SSTs. Related SSTs can be grouped into families, and these SST families may span multiple fold families. In terms of the description and analysis of protein architecture, SSTs possess many of the properties one would associate with important constituents of protein structure, making them a useful construct for creating a protein structure taxonomy.

Chapter 3 outlines my implementation of a nearest-neighbor classifier for the prediction of secondary structure. A key innovation is a novel similarity scoring scheme for identifying neighbors based on the local environment methodology developed by Bowie and Eisenberg (1991). The predictor was able to achieve a peak prediction accuracy of 68%. A second important idea is converting the final output into a probability distribution over the three secondary structure states instead of a one-state prediction. These probability estimates convey a sense of the reliability of each prediction. Finally, I have developed a measure of prediction performance based on the mutual information provided by the predictions about the true secondary structure.

Chapter 4 addresses the problem of fold prediction using a procedure that combines sequence-based and structure-based (local environment) information into a single, expanding template. The technique, termed iterative template refinement (ITR), employs an iterative search scheme that detects related proteins and sequentially adds them to the template. In this fashion, the initial seed template, constructed from a protein of known structure, gives rise to a tree of descendent templates containing both a structure-environment component as well as a multiple sequence alignment component. The enhanced 'signal' in these templates enable ITR to detect structural similarities between distantly related proteins.

Chapter 5 links together the methods, results, and conclusions from the middle three chapters and explores some of the broader implications of the thesis. Although on the surface each of the topics seem to represent an independent piece of research, upon closer inspection there are interesting interconnections. For example, the predictions from the nearest-neighbor secondary structure classifier could be used as input into the ITR method for fold recognition in place of the Bowie-Eisenberg scoring system. Similarly, the existence of SSTs shared by the Rossmann fold family and the TIM barrel family provides one possible explanation for the detection of p21 RAS and arabinose-binding protein by the tryptophan synthase (TIM barrel protein) ITR search. Finally, it is important to emphasize the relavance of this thesis to other areas of protein structure prediction and analysis including lattice simulations of protein folding, homology modelling, and protein design.

# References

[Altschul et al., 1990] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.*, 215:403–410.

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

[Barton and Sternberg, 1990] Barton, G. J. and Sternberg, M. J. E. (1990). Flexible protein sequence patterns: a sensitive method to detect weak structural similarities. *J. Mol. Biol.*, 212:389–402.

[Bleasby et al., 1994] Bleasby, A. J., Akrigg, D., and Attwood, T. K. (1994). OWL – a non-redundant composite protein sequence database. *Nucleic Acids Res.*, 22:3574–3577.

[Blundell and Johnson, 1993] Blundell, T. L. and Johnson, M. S. (1993). Catching a common fold. *Protein Sci.*, 2:877–883.

[Boresch et al., 1994] Boresch, S., Archontis, G., and Karplus, M. (1994). Free energy simulations: The meaning of the individual contributions from a component analysis. *Proteins Struct. Funct. Genet.*, 20:25–33.

[Bowie et al., 1991] Bowie, J. U., Luthy, R., and Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, 253:164–170.

[Carr and Kim, 1993] Carr, C. M. and Kim, P. S. (1993). A spring-loaded mechanism for the conformational change of influenza hemagglutinin. *Cell*, 73:823–832.

[Chan and Dill, 1991] Chan, H. S. and Dill, K. A. (1991). Polymer principles in protein structure and stability. *Annu. Rev. Biophys. Biophys. Chem.*, 20:447–490.

[Chothia, 1993] Chothia, C. (1993). One thousand families for the molecular biologist. *Nature*, 357:543–544.

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

[Chou and Fasman, 1974] Chou, P. Y. and Fasman, G. D. (1974). Prediction of protein conformation. *Biochemistry*, 13:222–244.

[Cohen et al., 1983] Cohen, F. E., Abarbanel, R. A., Kuntz, I. D., and Fletterick, R. J. (1983). A combinatorial approach to secondary structure prediction: $\alpha/\beta$ proteins. *Biochemistry*, 22:4894–4904.

[Creighton, 1989] Creighton, T. E., editor (1989). *Protein Structure: A Practical Approach*. IRL Press, Oxford.

[Creighton, 1993] Creighton, T. E. (1993). *Proteins: Structures and Molecular Properties*. W. H. Freeman and Company, New York.

[Daggett and Levitt, 1993] Daggett, V. and Levitt, M. (1993). Realistic simulations of native-protein dynamics in solution and beyond. *Annu. Rev. Biophys. Biophys. Chem.*, 22:353–380.

[Dill, 1990] Dill, K. A. (1990). Dominant forces in protein folding. *Biochemistry*, 29:7133–7155.

[Fersht, 1987] Fersht, A. R. (1987). The hydrogen bond in molecular recognition. *TIBS*, 12:301–304.

[Flores et al., 1994] Flores, T. P., Moss, D. S., and Thornton, J. M. (1994). An algorithm for automatically generating protein topology cartoons. *Protein Eng.*, 7:31–37.

[Gibrat et al., 1987] Gibrat, J. F., Garnier, J., and Robson, B. (1987). Further developments of protein secondary structure prediction using information theory. *J. Mol. Biol.*, 198:425–443.

[Godzik and Skolnick, 1992] Godzik, A. and Skolnick, J. (1992). Sequence-structure matching in globular proteins: Application to supersecondary and tertiary structure determination. *Proc. Natl. Acad. Sci. U.S.A.*, 89:12098–12102.

[Goldstein et al., 1992] Goldstein, R. A., Luthey-Schulten, Z. A., and Wolynes, P. G. (1992). Protein tertiary structure recognition using associative memory Hamiltonians with local interactions. *Proc. Natl. Acad. Sci. U.S.A.*, 89:9029–9033.

[Gribskov and Devereux, 1991] Gribskov, M. and Devereux, J., editors (1991). *Sequence Analysis Primer.* Stockton Press, New York.

[Hendrick and Hartl, 1993] Hendrick, J. P. and Hartl, F. U. (1993). Molecular chaperone functions of heat-shock proteins. *Annu. Rev. Biochem.*, 62:349–384.

[Holm and Sander, 1993] Holm, L. and Sander, C. (1993). Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.*, 233:123–138.

[Holm and Sander, 1994] Holm, L. and Sander, C. (1994). Parser for protein folding units. *Proteins Struct. Func. Genet.*, 19:256–268.

[Hughson et al., 1990] Hughson, F. M., Wright, P. E., and Baldwin, R. L. (1990). Structural characterization of a partly folded apomyoglobin intermediate. *Science*, 249:1544–1548.

[Jaenicke, 1991] Jaenicke, R. (1991). Protein folding: Local structures, domains, subunits, and assemblies. *Biochemistry*, 30:3147–3161.

[Jones et al., 1992] Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992). A new approach to protein fold recognition. *Nature*, 358:86–89.

[Kabsch and Sander, 1983a] Kabsch, W. and Sander, C. (1983a). Dictionary of protein secondary structures: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22:2577–2637.

[Kabsch and Sander, 1983b] Kabsch, W. and Sander, C. (1983b). How good are predictions of protein secondary structure? *FEBS Letters*, 155:179–182.

[Karlin et al., 1991] Karlin, S., Bucher, P., and Brendel, V. (1991). Statistical methods and insights for protein and DNA sequences. *Annu. Rev. Biophys. Biophys. Chem.*, 20:175–203.

[Karplus and McCammon, 1983] Karplus, M. and McCammon, J. A. (1983). Dynamics of proteins: Elements and function. *Annu. Rev. Biochem.*, 53:263–300.

[Kirschner and Bisswanger, 1976] Kirschner, K. and Bisswanger, H. (1976). Multifunctional proteins. *Annu. Rev. Biochem.*, 45:143–166.

[Koonin et al., 1994] Koonin, E. V., Bork, P., and Sander, C. (1994). Yeast chromosome III: New gene functions. *EMBO J.*, 13:493–503.

[Lee, 1992] Lee, C. (1992). Calculating binding energies. *Curr. Opin. Struct. Biol.*, 2:217–222.

[Lesk et al., 1989] Lesk, A. M., Branden, C., and Chothia, C. (1989). Structural principles of alpha/beta barrel proteins: the packing of the interior of the sheet. *Proteins Struct. Func. Genet.*, 5:139–148.

[Levin et al., 1986] Levin, J. M., Robson, B., and Garnier, J. (1986). An algorithm for secondary structure determination in proteins based on sequence similarity. *FEBS Letters*, 205:303–308.

[Levitt and Greer, 1977] Levitt, M. and Greer, J. (1977). Automatic identification of secondary structure in globular proteins. *J. Mol. Biol.*, 114:181–239.

[Lim and Sauer, 1989] Lim, W. A. and Sauer, R. T. (1989). Alternative packing arrangements in the hydrophobic core of lambda repressor. *Nature*, 339:31–36.

[Maiorov and Crippen, 1992] Maiorov, V. N. and Crippen, G. M. (1992). A contact potential that recognizes the correct folding of globular proteins. *J. Mol. Biol.*, 227:876–888.

[Matthews, 1975] Matthews, B. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, 405:442–451.

[Mitchell et al., 1989] Mitchell, E. M., Artymiuk, P. J., Rice, D. W., and Willett, P. (1989). Use of techniques derived from graph theory to compare secondary structure motifs in proteins. *J. Mol. Biol.*, 212:151–166.

[Murphy et al., 1990] Murphy, K. P., Privalov, P. L., and Gill, S. J. (1990). Common features of protein unfolding and dissolution of hydrophobic compounds. *Science*, 247:559–561.

[Neuwald and Green, 1994] Neuwald, A. F. and Green, P. (1994). Detecting patterns in protein sequences. *J. Mol. Biol.*, 239:698–712.

[Novotny et al., 1984] Novotny, J., Bruccoleri, R. E., and Karplus, M. (1984). An analysis of incorrectly folded protein models. Implications for structure predictions. *J. Mol. Biol.*, 177:787–818.

[Orengo, 1994] Orengo, C. A. (1994). Classification of protein folds. *Curr. Opin. Struct. Biol.*, 4:429–440.

[Orengo et al., 1992] Orengo, C. A., Brown, N. P., and Taylor, W. R. (1992). Fast structure alignment for protein databank searching. *Proteins Struct. Func. Genet.*, 14:139–167.

[Orengo et al., 1993] Orengo, C. A., Flores, T. P., Taylor, W. R., and Thornton, J. M. (1993). Identification and classification of protein fold families. *Protein Eng.*, 6:485–500.

[Ouzounis et al., 1993] Ouzounis, C., Sander, C., Scharf, M., and Schneider, R. (1993). Prediction of protein structure by evaluation of structure-sequence fitness:

Aligning sequences to contact profiles derived from three-dimensional structures. *J. Mol. Biol.*, 232:805–825.

[Pauling et al., 1951] Pauling, L., Corey, R. B., and Branson, H. R. (1951). The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain. *Proc. Natl. Acad. Sci. U.S.A.*, 37:205–211.

[Pearson and Lipman, 1988] Pearson, W. R. and Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA*, 85:2444–2448.

[Pickett et al., 1992] Pickett, S. D., Saqi, M. A. S., and Sternberg, M. J. E. (1992). Evaluation of the sequence template method for protein structure prediction. *J. Mol. Biol.*, 228:170–187.

[Qian and Sejnowski, 1988] Qian, N. and Sejnowski, T. J. (1988). Predicting the secondary structure of globular proteins using neural network models. *J. Mol. Biol.*, 202:865–884.

[Richardson, 1981] Richardson, J. S. (1981). The anatomy and taxonomy of protein structure. *Adv. Prot. Chem.*, 34:167–339.

[Rigby et al., 1986] Rigby, M., Smith, E. B., Wakeham, W. A., and Maitland, G. C. (1986). *The Forces between Molecules.* Clarendon Press, Oxford.

[Rost and Sander, 1993] Rost, B. and Sander, C. (1993). Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.*, 232:584–599.

[Sali et al., 1994] Sali, A., Shakhnovich, E., and Karplus, M. (1994). Kinetics of protein folding. A lattice model study of the requirements for folding to the native state. *J. Mol. Biol.*, 235:1614–1636.

[Sauer, 1993] Sauer, K., editor (1993). *Methods in Enzymology*, volume 246. Academic Press, Inc., San Diego.

[Schulz and Schirmer, 1979] Schulz, G. E. and Schirmer, R. H. (1979). *Principles of Protein Structure.* Springer-Verlag, New York.

[Serrano and Fersht, 1989] Serrano, L. and Fersht, A. R. (1989). Capping and alpha-helix stability. *Nature*, 342:296–299.

[Sharp and Honig, 1990] Sharp, K. A. and Honig, B. (1990). Electrostatic interactions in macromolecules: Theory and applications. *Annu. Rev. Biophys. Biophys. Chem.*, 19:301–332.

[Sippl and Weitckus, 1992] Sippl, M. J. and Weitckus, S. (1992). Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a database of known protein conformations. *Proteins Struct. Func. Genet.*, 13:258–271.

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

[Stout and Jensen, 1968] Stout, G. E. and Jensen, L. H. (1968). *X-ray Structure Determination: A Practical Guide.* Macmillan Publishing Co., Inc., New York.

[Taylor, 1986] Taylor, W. R. (1986). Identification of protein sequence homology by consensus template alignment. *J. Mol. Biol.*, 188:233–258.

[Taylor and Orengo, 1989] Taylor, W. R. and Orengo, C. A. (1989). Protein structure alignment. *J. Mol. Biol.*, 208:1–22.

[Tidor and Karplus, 1994] Tidor, B. and Karplus, M. (1994). The contribution of vibrational entropy to molecular association: The dimerization of insulin. *J. Mol. Biol.*, 238:405–414.

[Vita et al., 1989] Vita, C., Fontana, A., and Jaenicke, R. (1989). Folding of thermolysin fragments. *Eur. J. Biochem.*, 183:513–518.

[Wutrich, 1986] Wutrich, K. (1986). *NMR of Proteins and Nucleic Acids.* John Wiley & Sons, Inc., New York.

[Yee and Dill, 1993] Yee, D. P. and Dill, K. A. (1993). Families and the structural relatedness among globular proteins. *Protein Sci.*, 2:884–899.

[Yi and Lander, 1993] Yi, T. M. and Lander, E. S. (1993). Protein secondary structure prediction using nearest-neighbor methods. *J. Mol. Biol.*, 232:1117–1129.

[Zhang et al., 1992] Zhang, X., Mesirov, J. P., and Waltz, D. L. (1992). Hybrid system for protein secondary structure prediction. *J. Mol. Biol.*, 225:1049–1063.

# Chapter 2

# The Identification and Classification of Shared Structural Topologies (SSTs) at the Subdomain Level

## 2.1 Abstract

We have searched the database of known protein structures for shared structural topologies at the subdomain level. Our data consisted of 317 structural domains selected from the Brookhaven Databank. Using a fast and sensitive structural comparison technique, we performed an exhaustive pairwise comparison of all the domains. This work differs from previous attempts to classify protein structures in three important respects: (1) domains rather than complete polypeptide chains were used in the comparisons; (2) we have devised a novel measure of structural similarity normalized for alignment length; and (3) we have relaxed the requirement that the majority of positions in both structures are included in the alignment. Loosening the requirement for global alignment permitted the identification of common subdomains among structures considered to possess distinct structural folds. We refer to these recurring

recurring substructures as shared structural topologies (SSTs) as distinguished from the structural fold. We have clustered related SSTs into families and have demonstrated that members of different fold families may belong to the same SST family. From the standpoint of protein architecture, SSTs represent meaningful structural units that are useful for identifying relationships among different domains, pinpointing the core packing arrangment in domains, and uncovering component subdomains in complex structural folds. Finally, we demonstrate how SSTs can be used to analyze and describe a new protein structure using the recently solved structure of HIV reverse transcriptase as an example. We have discovered that the "palm" catalytic subdomain of HIV reverse transcriptase shares a common substructure with the carboxy-terminal domain of glutamine synthetase.

## 2.2 Introduction

Domains have long been considered a fundamental unit of protein structure, but in recent years, structural biologists have begun to study the properties of pieces of domains or subdomains. The subdomain level of structural organization has attracted attention for several reasons. First, the size of a typical subdomain is approximately 40 to 120 residues, which fills the considerable gap between the supersecondary structure level (approximately 20 to 40 residues) and the domain level (approximately 70 to 300 residues). Secondly, experimental evidence suggests that specific subdomains may be intermediates in the folding pathway (Hughson et al., 1990). Thirdly, certain subdomains form stable autonomous folding units when dissected away from the rest of the protein (Jaenicke, 1991). Finally, some subdomains in isolation display many of the properties of a molten globule (Peng & Kim, 1994).

One of the surprises in structural biology has been the observation that certain structural topologies (e.g., the four-helix bundle, TIM barrel, etc.) appear in a wide range of domains. These recurrent structural patterns at the domain level have been termed folds. Some have even speculated that the universe of folds may be quite limited – fewer than 1000 distinct folds (Chothia, 1992) – greatly simplifying the task of structural analysis and prediction. Following the lead of Jane Richardson, several groups have attempted a systematic classification of known structures into fold families (Holm & Sander, 1993; Orengo et al., 1994). There has been complementary research directed toward identifying structural patterns at the secondary and supersecondary structure levels (Ring et al., 1992; Rice et al., 1990). Surprisingly, this type of taxonomic analysis has not yet been applied to the subdomain level.

We have identified and classified recurrent structural patterns at the subdomain level by carrying out a systematic pairwise comparison of structural domains. Previous investigations have compared the known structures in the Brookhaven Data Bank with the goal of grouping structures into fold families (Orengo et al., 1993; Holm & Sander, 1993). Because we were interested in uncovering common topologies at the subdomain level, we adopted a more flexible strategy that incorporated three important methodological changes. First, our dataset was composed of protein domains rather than whole polypeptide chains. Secondly, we devised a measure of structural similarity normalized for alignment length so that a more precise threshold for significant structural similarity could be established. Finally and most importantly, we have relaxed the requirement for global alignment, thereby permitting the matching of common substructures.

The exhaustive pairwise comparisons successfully identified a large number of common substructures shared by two or more distinct domains. We refer to these recurring substructures as shared structural topologies (SSTs). SSTs spanned compact, globular regions of the protein, often encompassing interior residues involved in important packing interactions. SST families were created by grouping together related SSTs. Many SST families were quite large, including domains from multiple fold families. Indeed, the distinction between SSTs and structural folds was highlighted by the observation that the same SST could be found in structures possessing different overall folds.

Because they occupy a strategic position in the hierarchy of protein structural organization between the supersecondary structure and domain levels, SSTs have contributed important insights into protein architecture. From the perspective of the universe of possible folds, SSTs have proved to be a unifying influence, linking together seemingly unrelated structures into an extended family network. In terms of the analysis of individual structures, SSTs often coincide with the core packing region of a domain, thus pinpointing the critical packing arrangements. Finally, we have found that the larger and more complex domains may contain multiple SSTs that encompass different regions of the protein, thus highlighting different packing units (possibly overlapping) within the same structure. In the future, we propose that new structures be analyzed not only for folds, supersecondary structure motifs, and secondary structure elements, but also for SSTs. As a case study, we present a SST analysis of the HIV reverse transcriptase structure.

## 2.3 Materials and Methods

### (A) Dataset

We selected 226 representative protein chains from the July 1993 release of the Brookhaven database. We used the domain assignments supplied by the crystallographers to define the domains. The resulting 317 domains are listed in Table 1.

### (B) Calculating structural similarity

We wish to identify alignments between protein domains possessing similar structures. To make this goal precise, we must define the notion of an alignment and provide a measure of structural similarity between two aligned proteins. To this end, we have devised a normalized measure of structural similarity that is relatively independent of the alignment length.

**(a) Alignment.** Given domains $A$ and $B$ of respective lengths $n_A$ and $n_B$, an alignment of length L, $X_L^{AB}$, between them is defined to be the set of ordered pairs $\{(a_1, b_1), \cdots, (a_L, b_L)\}$ of corresponding positions in the two domains (with $1 \leq a_1 < a_2 < \cdots < a_L \leq n_A$ and $1 \leq b_1 < b_2 < \cdots < b_L \leq n_B$). Positions $a_i$ and $b_i$ are said to be corresponding positions.

**(b) Distance matrix.** Given domain $A$, the distance matrix consists of elements $d_{ij}^A$ defined to be the distance between the $i$-th and $j$-th alpha carbons of the protein.

**(c) DRMS between two structures.** Given an alignment $X_L^{AB}$ of length L between domains $A$ and $B$, the distance root-mean-square deviation (DRMS) of the alignment is defined to be

$$DRMS(X_L^{AB}) = \sqrt{\sum_{i<j}^{L} \sum_{j=1}^{L} \frac{(d_{a_i a_j}^A - d_{b_i b_j}^B)^2}{\frac{L \cdot (L-1)}{2}}}, \tag{2.1}$$

where $a_n$ and $b_n$ refer to the $n$-th aligned position in structures $A$ and $B$, respectively.

**(d) Finding "optimal" alignments.** To identify structural similarities, we would like to be able to find the optimal alignment with respect to DRMS. Although explicit optimization is expected to be time-consuming, we have developed a rapid heuristic algorithm which relies on iterative dynamic programming and bears some similarities to an algorithm of Subbiah et al. (1993). The details are given in Appendix 1.

**(e) Lower bound on alignment lengths.** To avoid uninteresting matches of short structural segments, it is useful to place a lower bound on the allowable length of the alignment. In the past, crystallographers have adopted the conservative stance of requiring that the alignments cover at least 60% of the larger protein. We have somewhat relaxed this criterion by defining $L_{min}(n)$, the minimum allowable alignment length for a domain of length $n$, to be the following piecewise linear function:

$$L_{min}(n) = \begin{cases} (0.25 \cdot n) + 25 & 60 \leq n < 100 \\ (0.2 \cdot n) + 30 & 100 \leq n < 150 \\ (0.4 \cdot n) & 150 \leq n < 250 \\ 100 & n \geq 250 \end{cases} \tag{2.2}$$

This criterion requires coverage of greater than 60% for domains of 60 residues, 50% for domains of 100 residues, 40% for domains of 150-250 residues, and 100 positions for domains exceeding 250 residues.

**(f) Normalized DRMS (nDRMS) and Distance Similarity Ratio (DSR).** DRMS itself turns out not to be a good measure for comparing the structural similarities among alignments of different lengths, because it is highly dependent on alignment length. Specifically, we generated pairs of random compact structures (as described in Appendix 2) and determined the best alignments between them (using the algorithm described in Appendix 1). The DRMS values scaled roughly linearly with alignment length (Figure 1). To eliminate the dependence on alignment length, we defined the normalized DRMS (nDRMS) of an alignment $X_L^{AB}$ of length $L$ by the equation

$$nDRMS(X_L^{AB}) = \frac{DRMS(X_L^{AB})}{random\_DRMS(L)}, \tag{2.3}$$

where $random\_DRMS(L)$ denotes the mean DRMS for alignments of length $L$ between random compact structures. Thus, the expected nDRMS score for random structures is 1. Finally, we defined $Y_L^{AB}$ to be the alignment of length $L$ with the minimum DRMS value.

For a given pair of structures $A$ and $B$, nDRMS is still a function of the alignment and exhibits some fluctuation for different alignment lengths (see Figure 2). We defined the Distance Similarity Ratio (DSR) to be the minimum nDRMS score for alignments that satisfy the minimum alignment length criterion:

$$DSR^{AB} = min(nDRMS(Y_L^{AB})), L_{min}(n_A) \leq L \leq n_s, \tag{2.4}$$

where $n_A$ is the length of structure $A$, and $n_s$ is the length of the smaller structure. In the work below, the DSR statistic was used to measure the degree of structural similarity for each comparison.

**(g) Definition of optimal and standard alignment.** We defined the alignment satisfying $L_{min}$ with the best nDRMS score to be the "optimal alignment". The length of the optimal alignment was invariably the minimum allowable length for the domains being compared. Often, there were other corresponding position pairs which when added to the alignment slightly increased the normalized DRMS value, but appeared to belong to the common substructure. Hence, we created the "standard alignment" by extending the optimal alignment to include all pairs of positions that satisfied the following criterion for a "good" position pair:

$$\text{single-}DRMS(X_L^{AB}, (a_i, b_i)) \leq random\_DRMS(L), \tag{2.5}$$

where the single position DRMS value (single-DRMS) for each position pair $(a_i, b_i)$ in the alignment was defined to be

$$\text{single-}DRMS(X_L^{AB}, (a_i, b_i)) = \sqrt{\sum_{j \neq i}^{L} \frac{(d^A_{a_i a_j} - d^B_{b_i b_j})^2}{L-1}}. \tag{2.6}$$

## (C) Definition of Shared Structural Topologies (SSTs)

Two structures $A$ and $B$ form a shared structural topology (SST) if $DSR^{AB} \leq 0.6$ ($SST^{AB}$) or $DSR^{BA} \leq 0.6$ ($SST^{BA}$). The SST is defined to be the standard alignment of the comparison. Note that the length of a SST is always greater than or equal to 40, but is not constrained to be less than or equal to 100 since the length of the standard alignment often exceeds $L_{min}$. More generally, it is possible to define a SST between more than two domains. A m-way SST is a substructure shared by $m$ distinct domains $\{(a_{11}, \cdots, a_{m1}), \cdots, (a_{1L}, \cdots, a_{mL})\}$.

## (D) Definition of SST Families

Each SST family was organized around a family center. Every domain in turn was selected as the center of a family, and the family members were determined according to degree of structural similarity to the family center. Thus, there were 317 family centers and 317 families. We explored two different criteria for family membership.

(1) *Single-center family.* All two-way SSTs that were derived from a family center were members of the family (i.e., $F_A = \{SST^{AM_1}, \cdots, SST^{AM_r}\}$ where $A$ is the family center and $M_i$ is another domain in the database).

(2) *Multiple-center family.* First, a set of close relatives to the family center were identified. $B_i$ was considered to be a close relative of domain $A$ if $DSR^{AB_i} \leq 0.5$ and if the standard alignment included at least 60% of the positions in both structures. The union of the single-center families of the center $A$ and its close relatives $B_i$ constituted the multiple-center family of $A$.

## (E) Choosing the N families with maximal coverage of the domain database

There was significant overlap between the various SST families, with many domains appearing in multiple families. We wished to choose the $N$ families that maximally covered the domain database. The total coverage of $N$ families was measured either (i) by the total number of different domains that contained a SST belonging to one of the families or (ii) by the sum of the fraction of positions in each domain that was included in a SST from one of the $N$ families. Starting with the largest SST family, we adopted the greedy strategy of successively adding the family that maximally increased the total coverage. This procedure is not guaranteed to find the optimal coverage, but should produce a reasonable approximation.

## (F) Constructing the minimum spanning tree

The results from the all-by-all comparisons could be schematically represented as a fully connected graph in which the nodes denote the domains and the edges represent the structural similarity between the domains measured in terms of a DSR score. The value of an edge between two nodes A and B was the minimum of $DSR^{AB}$ and $DSR^{BA}$. From this raw data, it was possible to generate a minimum spanning tree for any subset of the domain database using the Kruskal algorithm.

## 2.4 Results

### (A) Creation of domain database and evaluation of comparison method

Polypeptide chains in the July 1993 release of the Brookhaven database were checked for sequence similarity. Chains possessing greater than 30% sequence identity with another chain were removed. The 226 chains that remained after this filtering step were divided into domains based on the recommendations of the crystallographers. We note that this manual definition of domains agrees well with the automated domain assignments generated by the program of Holm and Sander (1994). There were 317 domains in all, which ranged in length from 59 to 450 residues.

The goal of structural comparison is to find the optimal alignment of positions between two structures that minimizes some measure of structural similarity. We have developed a method that uses iterative dynamic programming to minimize the r.m.s. deviation of alpha carbon distances (DRMS) over a range of alignment lengths (see Materials and Methods). The algorithm permitted gaps, but not the permutation or reversal of chain order. We wished to evaluate the accuracy of the method by comparing its alignments to those published in the literature or generated by other programs. We collected a set of 100 control alignments obtained from pairs of related structures (Orengo & Taylor, 1992). In 75 cases the alignments were nearly identical ($\geq$ 70% identical). In 10 cases the iterative dynamic programming procedure produced a better alignment as judged by the DRMS and RMS over the common residues. In 8 cases the alignments were different ($<$ 70% identical), but possessed roughly equivalent DRMS values, and in 7 cases the alignment was slightly worse. Thus, our structural comparison method produced alignments comparable in quality to other methods.

## (B) Measure of structural similarity normalized for alignment length and minimum alignment length criterion

Two popular measures of structural similarity are the root-mean-square deviation of alpha carbon positions after superposition (RMS) and the root-mean-square deviation of interresidue alpha carbon distances (DRMS). One deficiency with both RMS and DRMS is that they are quite sensitive to the exact number of aligned positions (Cohen & Sternberg, 1980). RMS and DRMS exhibit an approximately linear relationship to one another (Cohen & Sternberg, 1980; Maiorov & Crippen, 1994), and in this work, we have primarily used DRMS to assess structural similarity.

In order to explore the alignment length dependence of DRMS in more detail, we constructed a set of random polymer structures of length 200 that were compact and uniformly-packed as gauged by the radius of gyration and the number of alpha carbon contacts. The structures were generated using three different lattice models: (1) knight's walk lattice (24 basis vectors); (2) 20-vector lattice with a relative coordinate system; and (3) 20-vector relative lattice containing local conformation constraints (Appendix 2). The degree of structural similarity among the random chains was measured over a range of alignment lengths from 30 to 100 positions, and for each type of lattice, a plot of the average DRMS for approximately 20 comparisons versus the number of aligned positions is shown in Figure 1. The average of the three sets of data was well fit by a line: $y = 0.037L + 0.766$ ($r = 0.98$). Others have calculated the random DRMS for unrelated structures as a function of alignment length, but their comparison techniques did not permit gaps or insertions (Cohen & Sternberg, 1980; Alexandrov & Go, 1994).

Merging the DRMS value and the alignment length into a single function would greatly expedite the assessment of the significance of a structural comparison. To this end, we have devised the following "normalized" measure of structural similarity:

$$nDRMS(X_L^{AB}) = \frac{DRMS(X_L^{AB})}{random\_DRMS(L)}. \tag{2.3}$$

$X_L^{AB}$ is the alignment of length $L$ between domains $A$ and $B$, and $random\_DRMS(L)$, calculated using the fitted line in Figure 1, is the DRMS over $L$ aligned positions expected for two unrelated structures. We define $Y_L^{AB}$ to be the alignment of length $L$ with the minimum DRMS value. Plotting the variation in $nDRMS(Y_L)$ as a function of alignment length for a set of 1000 unrelated domain pairs (see Figure 2) demonstrates that the normalized DRMS is much less sensitive to alignment length than DRMS alone. As the number of equivalenced positions increased from 30% to 60% of $n_s$ (length of the smaller domain), the value of $\frac{nDRMS(Y_L)}{nDRMS(Y_{0.45 \cdot n_s})}$ remained remarkably constant.

Although the overall graph was relatively flat, for certain pairs of related structures, the value of $nDRMS(Y_L)$ increased dramatically for longer alignment lengths. In Figure 2, we have superimposed a plot of the ratio $\frac{nDRMS(Y_L)}{nDRMS(Y_{0.45 \cdot n_s})}$ for a comparison between the domains 1btc_2 and 1mns_1. There is a sharp rise in this graph because 1btc_2 and 1mns_1 share a common substructure, and when the alignment extends beyond this substructure, the $nDRMS(Y_L)$ score jumps significantly. We have investigated this phenomenon in more detail by calculating the number of significant similarities as determined by a $nDRMS(Y_L)$ score below 0.5 or 0.6 as a function of alignment length. As shown in Table 2, there was a five-fold increase in the number of significant similarities as $L$ decreased from 65 to 35% of $n_s$. These data underscore the point that new types of conserved topologies could be identified by reducing the number of aligned positions in a comparison.

One of the central differences between this work and previous investigations is the establishment of a less stringent minimum alignment length criterion. In the past, at least 60% of positions in both proteins were equivalenced in order to ensure global alignment of the structures (Orengo et al., 1992). We have adopted a more flexible approach in defining the minimum number of aligned positions in a structure of length $n$, $L_{min}(n)$. First, the range of $L_{min}$ was set between 40 and 100 residues, inclusive. These bounds correspond to the length of a typical subdomain. Within this range, $L_{min}$ was defined to be a piecewise linear function with respect to $n$ (see equation (4.6) in Materials and Methods). Since most domains are between 70 and 300 residues, $L_{min}$ typically ranged from 60% to 33% of the length of the domain. To verify the appropriateness of this criterion, we have examined the "standard alignment" length of a set of related pairs of proteins ($0.4 \leq nDRMS(Y_{0.5 \cdot n_s}) \leq 0.55$). The standard alignment is defined by extending the optimal alignment (alignment with best nDRMS score) to include all position pairs whose single-position DRMS value for the alignment is less than $random\_DRMS(L)$ (see Materials and Methods). As shown in Figure 3, $L_{min}$ nicely defines the lower bound of the distribution, showing that it does not exclude the smaller common substructures in related domains.

As depicted in Figure 2, the normalized DRMS scores still exhibit some fluctuation with respect to alignment length. Incorporating the $L_{min}$ function, which restricts the possible values of $L$, into the nDRMS function leads to the following measure of the structural similarity that depends on the domain length, not the alignment length:

$$DSR^{AB} = min(nDRMS(Y_L^{AB})), L_{min}(n_A) \leq L \leq n_s. \tag{2.4}$$

In order to better understand the properties of this measure, we studied the distribution of DSR scores for a set of 3600 dissimilar protein pairs. The list was derived by choosing pairs of domains that belonged to different structural classes ($\alpha$, $\beta$, $\alpha/\beta$ or $\alpha+\beta$). As shown in Figure 4, the plot hovered around 1.0 with a mean of 1.13. Thus, although estimated from artificial lattice chains, the $random\_DRMS(L)$ function reasonably approximates the level of structural similarity between unrelated real protein chains. Only a single comparison between the all-$\beta$ domain 1fc1a_1 and the $\alpha+\beta$ domain 1gd1o_2 had a DSR score of less than 0.6 (0.51). We then calculated the DSR scores for a set of 400 similar protein pairs in which the proteins in each pair belonged to the same structural family (Orengo et al., 1992). The mean DSR score for these comparisons was 0.47, and the scores ranged from 0.1 to 0.7. More importantly, a clear division between the two sets occurred at around 0.6. From the comparisons of the random lattice chains, we found that a DSR score of 0.6 was approximately 6 standard deviations below the mean of 1.0. Thus, based on the data from both the random lattice protein models as well as real protein chains, we have established a DSR score of 0.6 as the threshold for a significant structural similarity.

## (C) SSTs and SST families

Having assembled a database of protein domains, developed a method for comparing two protein structures, devised a normalized scale for structural similarity, and established a minimum alignment length criterion, we possessed the tools to perform an all-by-all comparison to identify common substructures. We refer to a substructure shared by two or more distinct domains to be a shared structural topology (SST). More precisely, two structures $A$ and $B$ form a SST if $DSR^{AB} \leq 0.6$ or $DSR^{BA} \leq 0.6$. The SST encompasses those positions that are in the standard alignment of the two structures. The set of aligned positions between the two related substructures specifies a "two-way" SST. More generally, the simultaneous alignment of $m$ related substructures gives rise to a m-way SST. Any SST can also be represented by a consensus distance matrix constructed by averaging the appropriate positions in the alpha carbon distance matrices of the involved structures.

Examination of the SSTs revealed that they form cohesive structural units and not loose amalgamations of secondary structure elements. As assessed by radius of gyration, SSTs are comparable in compactness to domains of the same size (within one standard deviation), and they contain a higher percentage of buried positions (37% versus 29%). The vast majority of SSTs ($> 85\%$) ranged in size from 40 (lower limit of $L_{min}$ function) to 120 residues, indicating that SSTs do indeed represent topologies at the subdomain level.

These findings were confirmed by direct visualization of some example SSTs. In Figure 5 we have used ribbons and topology diagrams as well as a direct superposition of alpha carbon atoms to display 2gbp_1 (glucose/galactose-binding protein, domain 1) and three SSTs derived from 2gbp_1. The region of each structure included in the SST is shaded. The three SSTs are characterized by a central $\beta$-sheet consisting of 4 or 5 parallel $\beta$-strands flanked by several helices. The shifting, addition, or removal of secondary structure elements has caused some interesting variation, but the overall topology of the SSTs is remarkably preserved. The SST encompasses the majority of positions in 5p21_1, whereas the SSTs with 2rus_2 and 1sbt_1 delineate a common subdomain within a larger structure. Finally, the conservation of the SSTs stands in contrast to the dissimilarity in the overall folds for the domains: 2gbp_1, (periplasmic-binding family) 2rus_2 (TIM barrel), 5p21_1 (Ras family) and 1sbt_1 (subtilisin serine protease family).

The vast number of SSTs generated by the comparisons were organized into families. We defined the SST families by selecting each domain in turn to serve as the center of a family and then recruiting those SSTs that were structurally related to this family center. Thus, there were 317 families and 317 family centers. In the simplest scheme, the family consisted of any SST derived from the family center (single-center family, see Figure 6A). A second, more generous definition of the family, identified a list of close relatives, $B_i$, to the family center $A$. Domain $B_i$ was considered to be a close relative of $A$ if $DSR^{AB_i} \leq 0.5$ and the standard alignment spanned at least 60% of the residues in both domains. The union of the single-center families of $A$ and its close relatives constituted the multiple-center family of $A$ (see Figure 6B).

Some of the SST families defined using the multiple-center method were quite large. The largest of the families were derived from $\alpha/\beta$ domains. For example, in Table 3, we list the members of the 2gbp_1 SST family. The family is composed of SSTs from 81 domains including members of a wide variety of fold families ranging from the TIM barrel, Rossmann fold, carboxypeptidase, glutathione reductase, and periplasmic-binding protein families. Most of these domains can be described as singly- or doubly-wound parallel $\beta$-sheet structures according to the Richardson taxonomy (Richardson, 1981). The shift from a single center to a multiple center added 16 domains to the family. In the work below, we exclusively used the multiple-center definition.

There was considerable overlap in membership among the 317 SST families. Several domains possessed SSTs belonging to many different families. In order to eliminate this redundancy, we have attempted to identify the 40 SST families maximally covering the domain database. Coverage was measured in terms of the total number of domains, number of domains with at least 50% of its positions included in a SST, and the sum of the fraction of each domain (fractional coverage) contained in a SST belonging to one of the families. We obtained the list in Table 4 using the greedy strategy of first choosing the largest SST family and then successively adding the family which increased the fractional coverage by the most. Both the individual coverage of each SST family as well as the running total coverage of the top 40 families is presented.

Surprisingly, the first three families in this list – 2lbp_2 (periplasmic-binding domain), 1tlk_1 (immunoglobulin fold), and 256b_1 (4-helix bundle) – contained SSTs from 141 domains. The top 40 families contained SSTs from over 241 domains or 75% of the database. Only 50 domains did not contain any SST at all, and many of these narrowly missed forming a SST (i.e., DSR score slightly above 0.6). By slightly relaxing the minimum alignment length criterion by 15% (i.e., $L_{min}^{0.85} = 0.85 \cdot L_{min}$), we reduced the number of domains without SSTs to 27. Thus, a disproportionate number of structures in the Brookhaven databank (July, 1993) possess a substructure belonging to one of a small number of SST families. It is likely that this trend will continue as more structures are added to the database. Furthermore, these data suggest that a new structure is likely to share a SST (not necessarily a previously identified SST) with an existing member of the domain database.

## (D) Protein architecture viewed from a SST perspective

Smaller than a structural fold and yet large enough to represent a defined packing unit, SSTs are an appropriate size to be considered a meaningful structural entity. In this role, SSTs have provided some important insights about protein structure.

It is possible to create a more unified description of a large set of structures by constructing a minimum spanning tree (MST) with SST links. The minimum spanning tree casts a broader net than the SST family since each node does not need to be connected to the center of the tree. We have constructed such a tree for the majority of $\alpha/\beta$ domains in our domain database (see Figure 7). The structure of the tree shows the relationship between a variety of $\alpha/\beta$ fold families. At the center of the MST are the periplasmic-binding protein domains (1dri_1, 2gbp_1, 2lbp_2, etc.). Emanating outward on one branch are the GTP-binding proteins (5p21_1 and 1etu_1) followed by the actin family (1atn_1, 2yhx_2, etc.). Radiating outward on another branch is the glutathione peroxidase family (1gp1a_1, 2trx_1, 1dsb_1). The large TIM barrel family (1tim_1, 1mns_2, 4xia_1, etc.) is also connected to the periplasmic-binding protein family. Down the trunk of the tree are the Rossmann fold domains (e.g., 8adh_2, 5ldh_2, 1gd1o_1, etc.) followed by 1dhr_1 and then the hydrolase family (4tgl_1, 1ace_1, 3sc2_1, 2eda_1). Branching off 8adh_2 is the glutathione reductase family (3grs_1, 3grs_2, 1phh_1, 1trb_1, etc.). The adenylate kinase family (3adk_1 and 1gky_1) and dihydrolate reductase (3dfr_1) are connected to the tree through 1dhr_1. Only three $\alpha/\beta$ domains (1alk_2, 1atr_1 and 2yhx_1) were not included in the MST.

Structural biologist often divide a protein structure into a core region essential for stability and outlying segments that are less crucial for the integrity of the protein. It is not trivial, however, to ascertain what constitutes the core, although visual inspection coupled with an analysis of residue packing and burial provides good hints. The fact that many SSTs are compact, well-buried, and conserved, suggests that they could furnish complementary information in this determination. In Figure 8, we present three examples of how aligning a set of SSTs derived from a given domain may help to define the structure's core. The first domain of the immunoglobulin FC fragment light chain, 1fc1a_1, shares SSTs with 15 other domains. In Figure 8B1, we have plotted the number of SSTs at each position in the structure. Positions with 14 or 15 SSTs were considered to constitute the core (see Figure 8A1), and this core corresponds quite closely to the framework residues identified by Chothia and Lesk (1987) in immunoglobulin structures. In the case of 3fbp_1, three SSTs confirm what is visually evident: the core of the structure corresponds to the middle 5 $\beta$-strands packed against the central helix (see Figure 8A2). Finally, staphylococcal nuclease possesses a single SST with 1bov_1 which is located not in the middle, but to one side of 2sns_1. This division of staphylococcal nuclease is supported by experimental data which shows that mutations in the twisted $\beta$-sheet spanned by the SST have a more profound effect on stability than mutations in the C-terminal helices (Meeker & Shortle, 1991).

The strict division of a structure into core and non-core regions is less enlightening for the larger domains which contain multiple important packing arrangments. SSTs, however, have also proved useful for describing these more complex folds in terms of component substructures (see Figure 9). For example, carboxypeptidase (5cpa_1) is a 307 residue $\alpha/\beta$ protein characterized by a long, mainly parallel $\beta$-sheet with helices on both sides. 5cpa_1 shares SSTs with 18 other domains, and many of these SSTs possess substantial overlap ($\geq$ 45% of positions in common) with one another. There are some exceptions, however. The SST with 1bia_2 shares the right half of the $\beta$-sheet, whereas the SST with 3sc2_1 encompasses the left half of the sheet (26% overlap). To complete the rough decomposition into a right part, left part, and central region, the SST with 2gbp_1 is located in the center of the sheet (38% overlap with the 1bia_2 SST and 34% overlap with the 3sc2_1 SST).

The globin fold has been extensively studied but an analysis using SSTs offers a new perspective on the structure. The SST between myoglobin (1mbd_1) and 2cro_1 delineates what can be considered the core packing arrangement between helices A, B, E, G, and H. On the other hand, the SST with 1hmq_1 shows that the two long helices G and H interact with one another in a manner similar to that found in helix 2 and helix 3 of four-helix bundle proteins. For the third example, we turn to two members of the TIM barrel family. The traditional, geometric description of the TIM barrel has treated the structure as a single, monolithic entity. Surprisingly, we found that 2gbp_1 makes a SST with the N-terminal half of the barrel in 1tim_1, whereas it forms a SST with the C-terminal half of the barrel in 2rus_2. Thus, these two SSTs divide the TIM barrel into two half barrels.

## (E) Analysis of a new protein structure using SSTs: HIV reverse transcriptase

Another potentially useful function for SSTs is to aid crystallographers in the description of new protein structures. Trying to digest a complex, unfamiliar structure into more recognizable pieces presents a formidable challenge. Initially, one might search for canonical folds, define the secondary structure elements, and look for supersecondary structure motifs or regular packing arrangements. Identifying SSTs would furnish complementary information. Below, we scanned the recently solved structure of HIV reverse transcriptase for SSTs.

HIV reverse transcriptase (HIV RT) is a 556 amino acid protein that catalyzes the synthesis of RNA from DNA. Kohlstaedt et al. (1992) solved the structure to 3.5 Å and defined two distinct domains: a polymerase domain and a RNase H domain. We subjected the polymerase domain to SST analysis by comparing it to each of the 317 domains in our database. There were two significant hits: the region from residue 91 to 235 formed a SST with the N-terminal domain of RUBISCO (2rus_1) and the region from residue 322 to 415 formed a SST with *E. coli* RNase H (1rnh_1). These two SSTs divided in an approximate fashion the polymerase domain into the four parts which correspond to the four subdomains described by Steitz and collegues (Kohlstaedt et al., 1992).

Each of the four subdomains were then run against the domain database to search for more SSTs. The results are presented in Table 5. The "finger" subdomain did not register a hit to any of the domains. Likewise, the "thumb" subdomain exhibitted only weak structural similarity to several four-helix bundle proteins. The modest DSR scores support the assertion by Kohlstaedt et al. (1992) that the thumb subdomain distantly resembles a four-helix bundle architecture. As described above, the section of the protein from 320 to 420, termed the "connection" subdomain, shares a SST with 1rnh_1, and it also records significant DSR scores against three other members of the 1rnh_1 SST family, 1lap_2, 1atn_1, and 1atn_2. This resemblance was previously noted by Artymiuk et al. (1993). In Figure 10A, we present a superposition of the connection sudomain against the RNase H domain of HIV RT. The close structural relationship between these two adjacent sections of the protein suggests a duplication event during the evolution of the HIV RT gene.

The most surprising finding was that the "palm" subdomain, encompassing positions 80 to 115 and 150 to 240 which includes the active site of the enzyme, contains an unexpected SST with the C-terminal domain of glutamine synthetase (2gls_2). The structural similarity is quite good: 2.22 Å RMS over 67 residues (see Figure 10B). Joyce and Steitz (1994) have speculated that three active site aspartic acid residues – D110, D185, and D186 – are involved in coordinating two $Mg^{2+}$ ions. These divalent metal ions are thought to catalyze the formation of the new phosphodiester bond. Intriguingly, in the structural alignment of 2gls_2 with the palm subdomain, $D110^{RT}$ is superimposed on $E131^{GS}$ and $D186^{RT}$ matches up with $E220^{GS}$. Both $E131^{GS}$ and $E220^{GS}$ are known to coordinate $Mn^{2+}$ in the active site of glutamine synthetase, lending indirect support for the Joyce-Steitz model. To our knowledge, this is the first time the catalytic subdomain of HIV RT has been shown to possess structural similarity with another non-polymerase protein.

## 2.5 Discussion

In this work, we have identified recurrent structural patterns at the subdomain level by completing an all-by-all structural comparison of 317 domains selected from the Protein Data Bank. This study was made possible by three important technical innovations. First, we have developed a fast and sensitive structural comparison method. Secondly, we were able to assess the significance of the comparisons by using a measure of structural similarity (DSR) normalized for alignment length. Finally, we have devised a minimum alignment length criterion that reduced the required number of equivalenced positions between the two structures, thereby facillitating the identification of smaller substructures than in previous studies. Making $L_{min}$ more stringent would result in the identification of related folds, whereas making $L_{min}$ less stringent would delineate common supersecondary structures.

We refer to these common substructures as common structural topologies or SSTs. The typical SST ranges from 40 to 120 residues which in some cases is large enough to span a domain, but in most cases represents only part of a domain. SSTs are compact and contain a higher proportion of buried residues. In the hierarchy of protein structural organization, SSTs fill an important gap between the structural fold and the supersecondary structure motif (see Figure 11). The archetypal structural patterns associated with each level greatly simplifies the analysis of a protein by assisting in the deconstruction of a complex structure into more familiar pieces. This theoretical work on common topologies at the subdomain level complements the ongoing experimental work elucidating the fundamental role of subdomains in protein folding.

We have organized the vast number of SSTs into families by clustering together related SSTs. One key finding of this study is that SST families are quite large, and domains from different fold families may contain SSTs belonging to the same SST family. For example, the 2gbp_1 SST family includes members of the TIM barrel, Rossmann fold, glutathione reductase, carboxypeptidase, and GTP-binding families. Altogether, the three largest non-overlapping SST families include 141 domains or 44% of the database. Only 50 domains possess no SSTs, and if the minimum alignment length requirement is relaxed slightly, this number is lowered to 27. In the near future as the structural database grows, it may be possible to describe a good proportion of every structure by referring to its SSTs.

In this work, we have argued that SSTs are interesting and important structural components. They are small enough to represent a single, relatively simple packing arrangment, and yet they are large enough to be considered a whole packing unit. We have shown that in this role SSTs have provided important insights into protein structure, showing the underlying commonality among different domains, helping to identify the core regions of a structure, and furnishing a description of more complicated folds in terms of component (possibly overlapping) substructures. All in all, SSTs provide the basis for a rich new taxonomy for the understanding of protein structure. In the future, we recommend that a new structure be described at both the fold level and the SST level, thereby furnishing a more complete picture of the structure.

Recently, Yee and Dill (1992) have argued that fold families as traditionally defined are loosely-knit forms of organization. In other words, the relationships within a family are not that much stronger than the relationships between different families. Their conclusions, however, were weakened by the fact that their comparison algorithm did not permit gaps or insertions, casting some doubt on the appropriateness of their measure of structural similarity (Orengo, 1994). Our work with SSTs, which uses a structural comparison technique that allows gaps and insertions, supports their basic claim since we have demonstrated that it is not uncommon for members of different fold families to share a common substructure. Thus, we agree with the assertion of Yee and Dill that the boundaries between different fold families are difficult to establish.

Within the structural biology community there has been a preoccupation with the enumeration and cataloguing of structural folds. This line of investigation has proved most valuable since there is considerable evidence that the number of structural folds is limited. On the other hand, this work has pointed out some limitations of the "fold-centric" view of the world. First, as mentioned above, an exclusive focus at the fold embellishment of a common SST with extra secondary and supersecondary structure elements may give rise to a spectrum of different folds. Secondly, there is the tendency to treat structural folds such as the TIM barrel as an inviolate whole, rather than perceiving the possibility of breaking up the structure into simpler pieces. Finally, from the perspective of protein design, one must appreciate the exciting prospect of mixing and matching SSTs to create new folds.

In an attempt to determine the basic unit of protein folding, experimentalists have begun to dissect proteins into smaller domains and subdomains and have asked whether these fragments behave as autonomous folding units (Jaenicke, 1991). An autonomous folding unit is operationally defined as a section of a protein that is stable in isolation. Many SSTs represent relatively self-contained subdomains that can be distinguished by limited contacts with the surrounding protein. Do some SSTs correpond to autonomous folding units? Only a handful of experiments have been performed to measure the stability of isolated subdomains. Vita et al. (1989) have made fragments of thermolysin and found that one of the smallest pieces that was compact and monomeric corresponded to residues 225 to 316 in the second domain of the protein. Interestingly, a weak SST between 3tln_2 and 2lzm_2 ($\mathrm{DSR} = 0.615, \mathrm{DSR}^{0.85} = 0.55$) encompassed the positions 171 - 179, 233 - 246, 257 - 272, 279 - 290, and 303 - 311 in thermolysin or basically the same region as the stable subdomain.

Why are these topologies conserved? Is it the consequence of divergent or convergent evolution? An examination of the conformation and positioning of the side chains in the SSTs demonstrates that the structural conservation is not only at the level of the alpha carbon atoms but also the at the level of the packing of the side chains. Indeed, there is a strong correlation in the atomic level residue-residue contact maps derived from the related substructures in a SST. The absence of any sequence similarity among proteins related by a shared SST and the fact that the same SST can appear in domains with significantly different overall structures suggest that physicochemical constraints, and not evolutionary descent, is responsible for the repeated use of the same architecture. In the future, we hope to study in more detail the factors that may explain the preference of certain packing arrangments.

Finally, we propose that every new structure be searched for SSTs along with recurrent structural patterns at other levels of structural organization. We have identified several interesting SSTs in the polymerase domain of HIV reverse transcriptase. First, we demonstrated that the connection subdomain is structurally similar to the RNAse H domain of RT. Secondly, our analysis has revealed that the palm subdomain shares a SST with the second domain of glutamine synthetase. There is a suggestive conservation of catalytic carboxylates in the two active sites. In the future, we hope to examine in more detail the evolutionary relationship between these two enyzmes.

## 2.6 Appendix 1: Structural Comparison Algorithm

We employed an iterative technique based on dynamic programming to find the optimal alignment. Unlike the the method of Holm and Sander (1993), this algorithm does not permit the reversal of chain direction and enforces the topological connectivity of the aligned segments (i.e., $1 \leq a_1 < \cdots < a_L \leq n_A$ and $1 \leq b_1 < \cdots < b_L \leq n_B$). A second important point is that the aligned segments were required to be at least three residues long to prevent the matching of short one or two amino acid fragments.

The basic idea is that the alignment at iteration $k$, $X^{AB,k}$, was used to calculate the cost matrix $M^k$. The value of $M^k_{ij}$ was inversely related to the root-mean-square deviation of the entries in row $i$ (belonging to the alignment) of the distance matrix of protein $A$ with respect to the corresponding entries in row $j$ of the distance matrix of protein $B$:

$$M^k_{ij} = (\frac{1}{\sqrt{\sum_m^L \frac{(d^A_{ia_m} - d^B_{jb_m})^2}{L}}} - \theta). \qquad (2.7)$$

The Needleman-Wunsch 'global' dynamic programming algorithm (Needleman & Wunsch, 1970) with no gap penalties produced a new alignment, $X^{AB,(k+1)}$, which was used to fill $M^{(k+1)}$. The whole process was repeated and the alignment usually converged in fewer than 8 iterations. The parameter $\theta$ controlled the length of the alignment and was lowered from 0.25 to 0.125 (i.e., $\theta = \frac{1}{4}, \frac{1}{5}, \frac{1}{6}, \frac{1}{7}, \text{and} \frac{1}{8}$; with 8 iterations for each value of $\theta$) in order to extend the alignment. Aligned segments of fewer than three amino acids were removed from the alignment. For each alignment $X^{AB,k}_L$, the DRMS values for a range of alignment lengths between $L_{min}$ and $L$ were calculated. An alignment of length L could be pruned to give rise to a set of shorter alignments by incrementally removing those position pairs with the worst single-DRMS values. The best DRMS score for each alignment length was stored.

It should be noted that this approach is somewhat similar to the independently derived iterative dynamic programming method described by Subbiah et al. (1993). The main difference is that Subbiah calculated the cost matrix $M^k$ by superimposing structures $A$ and $B$ given the alignment at iteration $k$ and then determined the absolute distance between residue $i$ of protein $A$ and residue $j$ of protein $B$.

Two methods were used to generate the initial alignment. First, we adapted the graph matching algorithm described by Grindley et al. (1993) to align secondary structure elements. Briefly, the protein structure was represented as a labelled graph in which the nodes correspond to secondary structure elements and the edges represent distance and angle relationships between these elements. A maximal common subgraph isomorphism algorithm identified similar structural patterns in the two structures and aligned equivalent secondary structure elements. To generate the initial alignment of individual positions, the middle residue of equivalent secondary structure elements were superimposed, and the alignment was extended in both directions to the boundaries of the elements.

Secondly, we implemented the more laborious approach of systematically aligning the positions of the two domains at different registers (i.e., $X = \{(1,1), \cdots, (n_A, n_A)\}$, offset $= 0$). Not all possible registers were tried, only those that were 10 positions apart (i.e., offset $= \cdots, -20, -10, 0, 10, 20, \cdots$). The iterative dynamic programming algorithm described above refined each of these initial alignments, and the best DSR score from the various offsets was recorded.

For most comparisons, the two methods produced the same final alignment. Method 1 was much faster and was used for the comparisons between real protein domains. Method 2 was used to compare the random lattice structures because they often did not possess true secondary structure.

Finally, after all the comparisons had been performed, we checked the consistency of the various alignments. First, we examined the agreement between $X^{AB}$ and $X^{BA}$. In other words, the same set of corresponding position pairs in alignment $X^{AB}$ should be found in the alignment $X^{BA}$ with only the order of the pair reversed (inverse relationship). Then, we also checked that $X^{AB}$ and $X^{BC}$ were consistent with $X^{AC}$. The three alignments were consistent at aligned position $i$ if given the three corresponding position pairs $(a_i, b_i) \in X^{AB}$, $(b_i, c_i) \in X^{BC}$, and $(a_i, c_i') \in X^{AC}$, $c_i' = c_i$. Any inconsistencies were resolved by checking if one or more of the alignments were not optimal.

## 2.7 Appendix 2: Generating Random Lattice Structures

Three different types of lattices were used to generate the random structures: (1) knight's walk lattice consisting of 24 basis vectors (Skolnick et al., 1990); (2) 20-vector lattice consisting of 20 basis vectors in a relative coordinate system (i.e., the $xyz$ axes at position $i$ were determined by the location of alpha carbon atoms $(i-2)$, $(i-1)$, and $i$; and (3) 20-vector lattice with local conformation constraints to encourage the formation of secondary structure. The 20-vector 'relative' lattice was constructed by selecting and adjusting 20 of the 24 basis vectors from the knight's walk lattice according to how well they could represent real protein structures in a relative coordinate system. Random structures with local secondary structure were created by fitting a real structure onto the lattice and then randomizing the vectors at coil positions. The secondary structure assignments were taken from the first 200 residues of the following structures: 1tim (46% $\alpha$, 17% $\beta$), 2cpp (49% $\alpha$, 6.5% $\beta$), 2cyp (42.5% $\alpha$, 2% $\beta$), and 3cna (0% $\alpha$, 45% $\beta$).

To ensure uniform packing, the radius of gyration and the number of alpha carbon contacts at 4 Å and 8 Å for each random structure were constrained to be within 2 standard deviations of the values expected for a real protein of the same size ($15.0 \leq R_G \leq 18.5$, $CONTACT_{4.0} \leq 20$, and $315 \leq CONTACT_{8.0} \leq 479$). We ran simulations allowing the lattice structures to sample conformational space under the influence of an energy function that measured the deviation from the expected values for the radius of gyration and the two types of contacts. Typically, over 10,000 iterations of simulated annealing were necessary to generate each compact random structure.

For the lattices without local conformational constraints, (1) and (2), two groups of five random structures were created. All possible comparisons within each group of five were calculated, resulting in $(2 \cdot 10) = 20$ total comparisons. For the third type of lattice, two random structures were created for each of the four secondary structure assignments. Comparisons between structures with the same local constraints produced DRMS values that were approximately 10% lower than comparisons between structures with different local structure. Only the data from the comparisons between structures with different local conformational constraints ($4 \cdot 6 = 24$ in total) were included in the calculation of the $random\_DRMS$ function.

# References

[Alexandrov and Go, 1994] Alexandrov, N. N. and Go, N. (1994). Biological meaning, statistical significance, and classification of local spatial similarities in nonhomologous proteins. *Protein Sci.*, 3:866–875.

[Artymiuk et al., 1993] Artymiuk, P. J., Grindley, H. M., Kumar, K., Rice, D. W., and Willett, P. (1993). Three-dimensional structural resemblance between the ribonuclease H and connection domain of HIV reverse transcriptase and the ATPase fold revealed using graph theoretical techniques. *FEBS Lett.*, 324:15–21.

[Branden and Tooze, 1991] Branden, C. and Tooze, J. (1991). *Introduction to Protein Structure.* Garland Publishers Inc., New York.

[Chothia, 1992] Chothia, C. (1992). One thousand families for the molecular biologist. *Nature*, 357:543–544.

[Chothia and Lesk, 1987] Chothia, C. and Lesk, A. M. (1987). Canonical structures for the hypervariable regions of immunoglobulins. *J. Mol. Biol.*, 196:901–917.

[Cohen and Sternberg, 1980] Cohen, F. E. and Sternberg, M. J. E. (1980). On the prediction of protein structure: The significance of the root-mean-square deviation. *J. Mol. Biol.*, 138:321–333.

[Grindley et al., 1993] Grindley, H. M., Artymiuk, P. J., Rice, D. W., and Willett, P. (1993). Identification of tertiary structure resemblance in proteins using a maximal common subgraph isomorphism algorithm. *J. Mol. Biol.*, 229:707–721.

[Holm and Sander, 1993] Holm, L. and Sander, C. (1993). Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.*, 233:123–138.

[Holm and Sander, 1994] Holm, L. and Sander, C. (1994). Parser for protein folding units. *Proteins Struct. Func. Genet.*, 19:256–268.

[Hughson et al., 1990] Hughson, F. M., Wright, P. W., and Baldwin, R. L. (1990). Structural characterization of a partly folded apomyoglobin intermediate. *Science*, 249:1544–1548.

[Jaenicke, 1991] Jaenicke, R. (1991). Protein folding: local structures, domains, subunits, and assemblies. *Biochemistry*, 30:3147–3161.

[Jones et al., 1992] Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992). A new approach to protein fold recognition. *Nature*, 358:86–89.

[Klug and Rhodes, 1987] Klug, A. and Rhodes, D. (1987). Zinc fingers: A novel protein fold for nucleic acid recognition. *Cold Spring Harbor Symp. Quant. Biol.*, 52:473–482.

[Kohlstaedt et al., 1992] Kohlstaedt, L. A., Wang, J., Friedman, J. M., Rice, P. A., and Steitz, T. A. (1992). Crystal structure at 3.5 Å resolution of HIV-1 reverse transcriptase complexed with an inhibitor. *Science*, 256:1783–1790.

[Maiorov and Crippen, 1994] Maiorov, V. N. and Crippen, G. M. (1994). Significance of root-mean-square deviation in comparing three-dimensional structures of globular proteins. *J. Mol. Biol.*, 235:625–634.

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

[Orengo et al., 1993] Orengo, C., Flores, T. P., Taylor, W. R., and Thornton, J. M. (1993). Identification and classification of protein fold families. *Protein Eng.*, 6:485–500.

[Orengo, 1994] Orengo, C. A. (1994). Classification of protein folds. *Curr. Opin. Struct. Biol.*, 4:429–440.

[Orengo et al., 1994] Orengo, C. A., Jones, D. T., and Thornton, J. A. (1994). Protein superfamilies and domain superfolds. *Nature*, 372:631–634.

[Peng and Kim, 1994] Peng, Z. Y. and Kim, P. S. (1994). A protein dissection study of a molten globule. *Biochemistry*, 33:2136–2141.

[Rice et al., 1990] Rice, P. A., Goldman, A., and Steitz, T. A. (1990). A helix-turn-strand structural motif common in alpha-beta proteins. *Proteins Struct. Funct. Genet.*, 8:334–340.

[Richardson, 1981] Richardson, J. (1981). The anatomy and taxonomy of protein structures. *Adv. Protein Chem.*, 34:167–337.

[Ring et al., 1992] Ring, C. S., Kneller, D. G., Langridge, R., and Cohen, F. E. (1992). Taxonomy and conformational analysis of loops in proteins. *J. Mol. Biol.*, 224:685–699.

[Sauer et al., 1982] Sauer, R. T., Yocum, R. R., Doolittle, R. F., Lewis, M., and Pabo, C. O. (1982). Homology among DNA-binding proteins suggests use of a conserved super-secondary structure. *Nature*, 298:447–451.

[Schulz and Schirmer, 1979] Schulz, G. E. and Schirmer, R. H. (1979). *Principles of Protein Structure.* Springer-Verlag, New York.

[Serrano et al., 1991] Serrano, L., Matouschek, A., and Fersht, A. R. (1991). The folding of an enzyme. VI. the folding pathway of barnase: Comparison with theoretical models. *J. Mol. Biol.*, 224:847–859.

[Skolnick and Kolinski, 1990] Skolnick, J. and Kolinski, A. (1990). Simulations of the folding of a globular protein. *Science*, 250:1121–1125.

[Subbiah et al., 1993] Subbiah, S., Laurents, D. V., and Levitt, M. (1993). Structural similarity of DNA-binding domains of bacteriophage repressors and the globin core. *Curr. Biol.*, 3:141–148.

[Vita et al., 1989] Vita, C., Fontana, A., and Jaenicke, R. (1989). Folding of thermolysin fragments. *Eur. J. Biochem.*, 183:513–518.

[Yee and Dill, 1993] Yee, D. P. and Dill, K. A. (1993). Families and the structural relatedness among globular proteins. *Protein Sci.*, 2:884–899.

[Yi and Lander, 1994] Yi, T. M. and Lander, E. S. (1994). Recognition of related proteins by iterative template refinement (ITR). *Protein Sci.*, 3:1315–1328.

## 2.8 Figures and Tables

## Figure 1

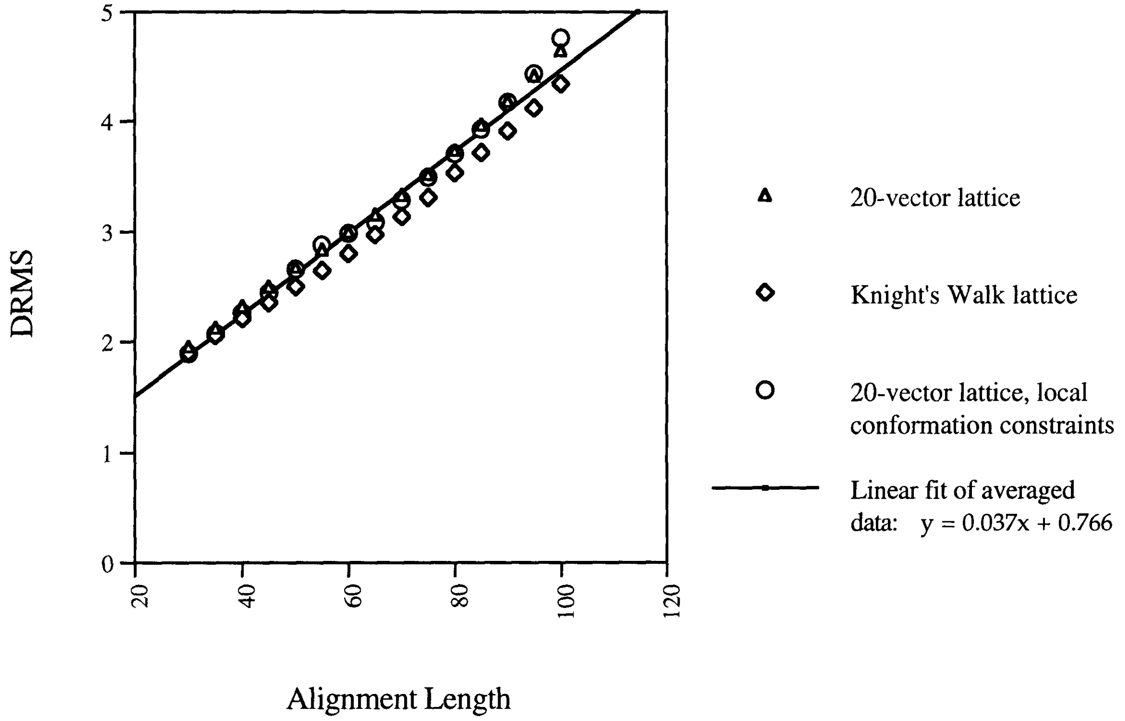


**Figure 1.** Plot of random DRMS versus alignment length. The random DRMS values were generated from the comparisons of random lattice polymer structures over a range of alignment lengths. Three types of lattice models were used, and each data point represents the average of 20 comparisons for a particular lattice. The equation of the line that best fit the average of the three sets of data is given.

## Figure 2

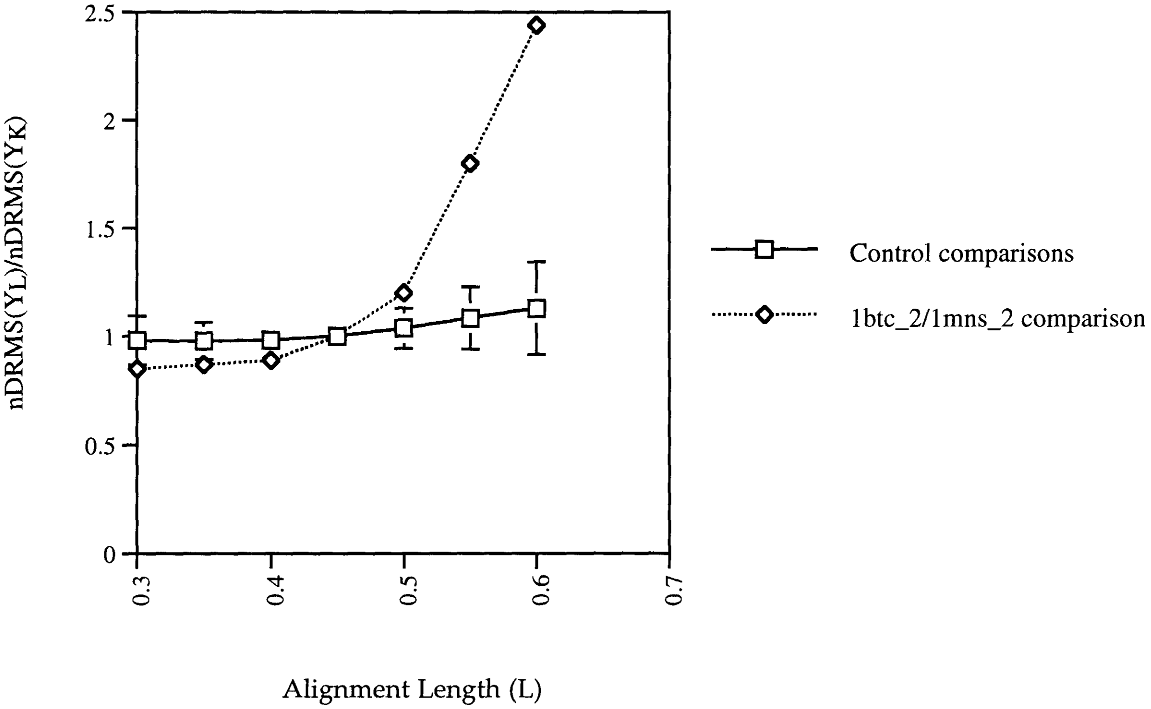


**Figure 2.** Variation in nDRMS as a function of alignment length. The nDRMS values are plotted as a ratio of nDRMS($Y_L$) to nDRMS($Y_K$) where $Y_L$ is the alignment of length L with the minimum DRMS and K = (0.45 • $n_s$). Alignment length is measured as a fraction of the length of the smaller domain ($n_s$). The square data points represent the average of 1000 control comparisons, whereas the diamond data points are the data from a single comparison between the domains 1btc_2 and 1mns_2.

# Figure 3

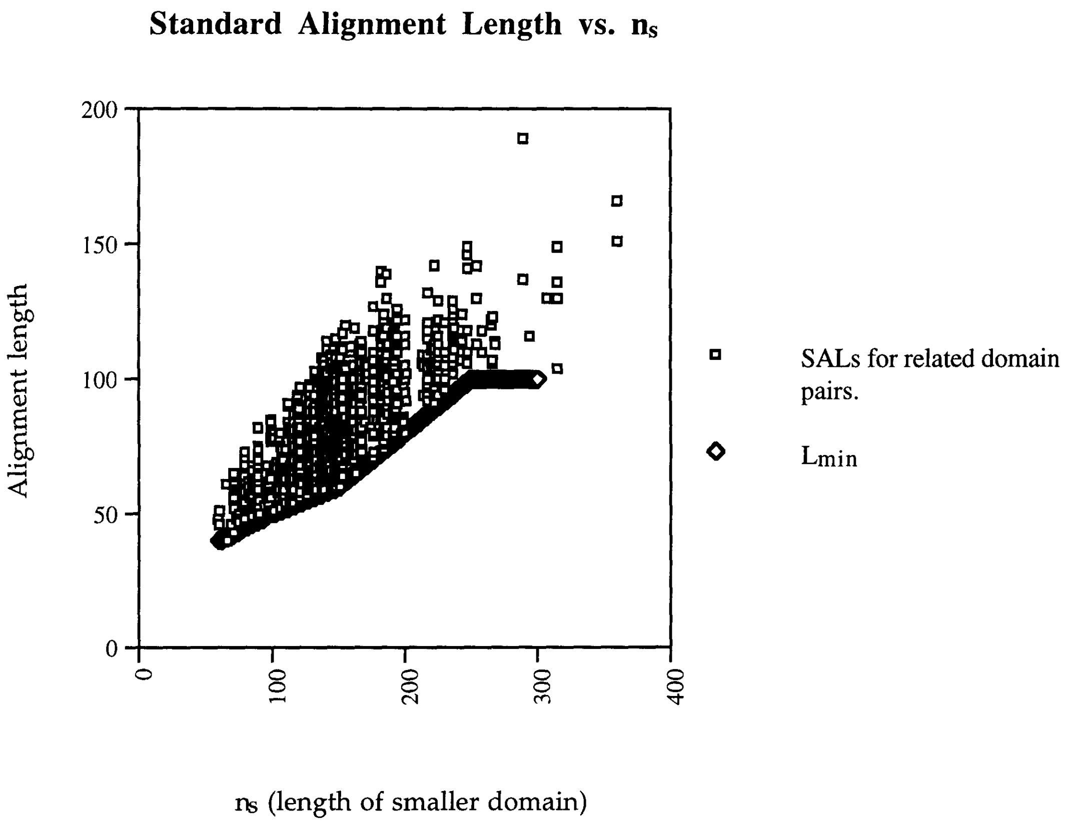


**Figure 3.** Distribution of standard alignment lengths (SALs) as a function of domain length. The alignments were generated from a set of related domain pairs. Superimposed is a graph of the minimum alignment length function, $L_{min}$.

## Figure 4

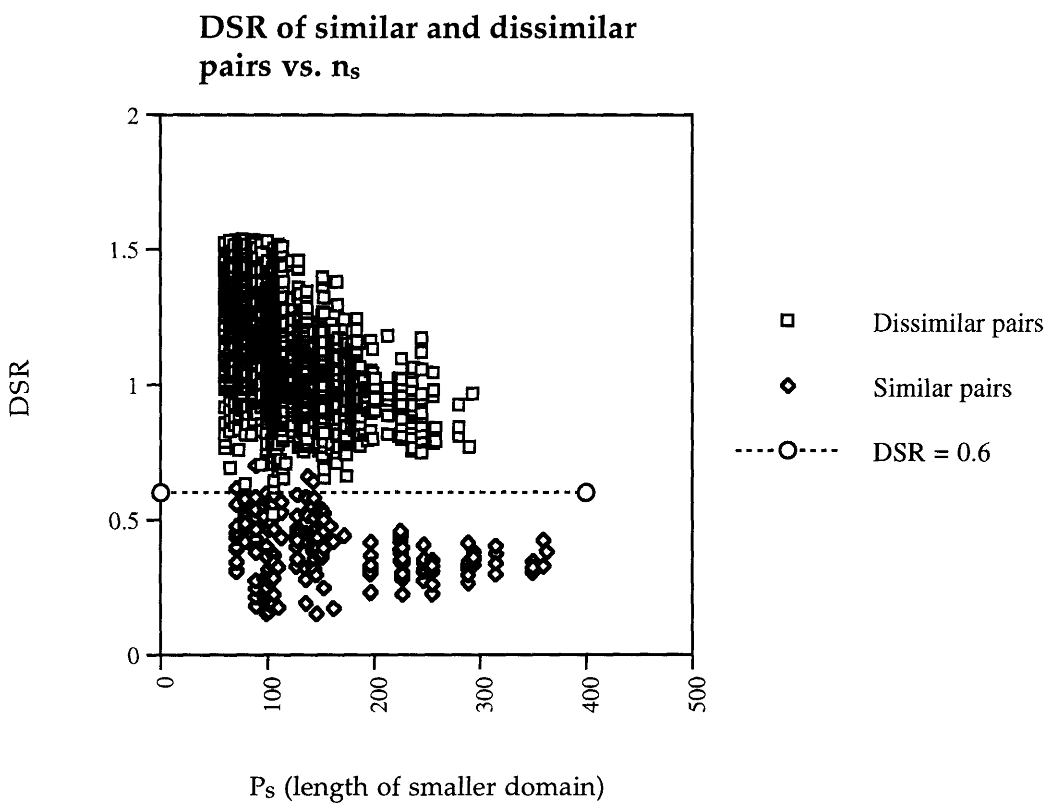
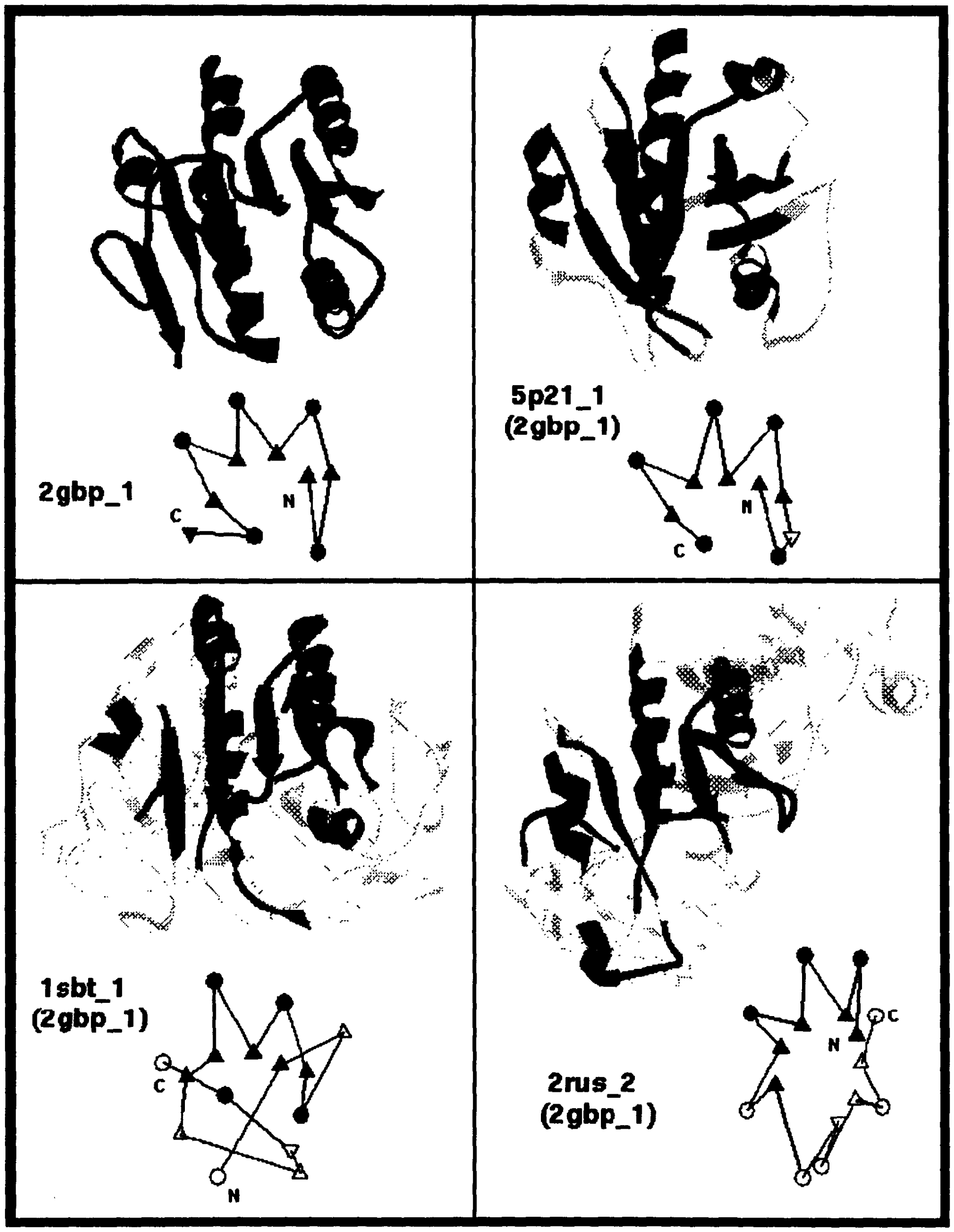
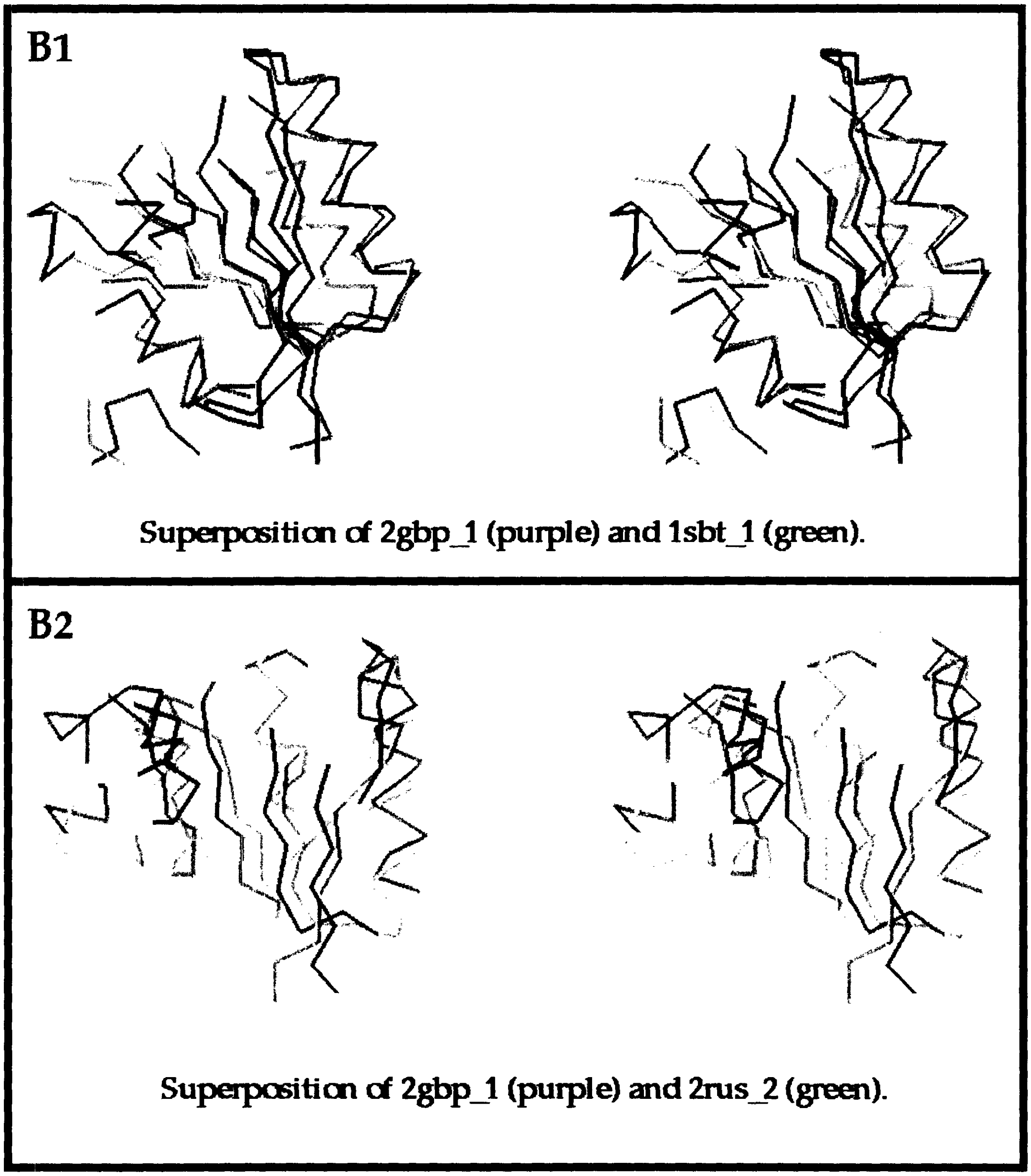


**Figure 4.** Distribution of DSR values for dissimilar pairs (squares) and similar pairs (diamonds). The dotted line signifies the value of DSR = 0.6, which is the cutoff for a significant structural similarity.

# Figure 5A



# Figure 5B



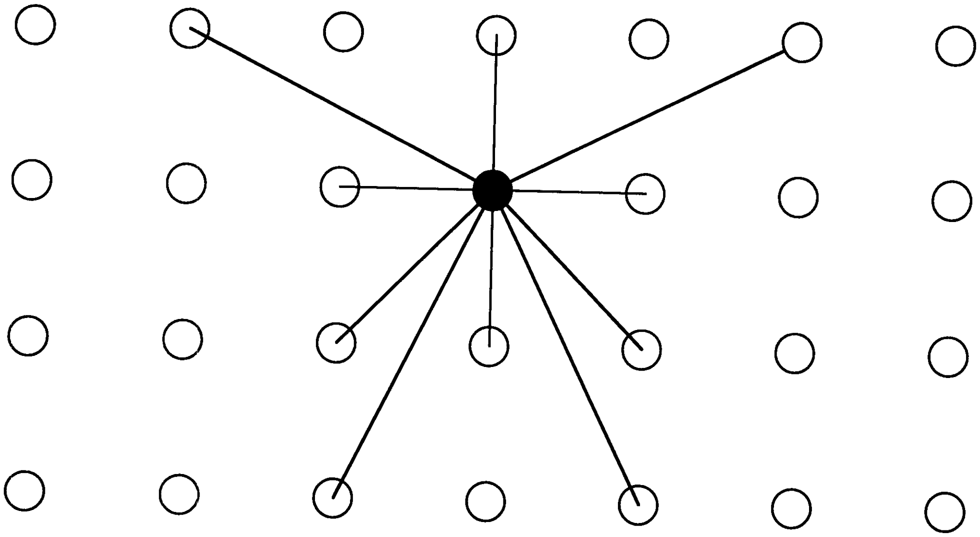
Superposition of 2gbp_1 (purple) and 1sbt_1 (green).

Superposition of 2gbp_1 (purple) and 2rus_2 (green).

**Figure 5.** Examples of SSTs derived from 2gbp_1. (A) Ribbon and topology diagrams of the domain 2gbp_1 and 3 other domains sharing a SST with 2gbp_1: (1) 2gbp_1, (2) 5p21_1, (3) 1sbt_1, (4) 2rus_2. The SST is shaded in purple in the ribbon diagrams and depicted by the filled circles (helices) and triangles ($\beta$-strands) in the topology diagrams. (B) The alpha carbon superposition of related substructure between (1) 2gbp_1 (purple) and 1sbt_1 (green), and (2) 2gbp_1 (purple) and 2rus_2 (green). Only positions present in the SST are shown.

## Figure 6

(A) Single Center



(B) Multiple Center

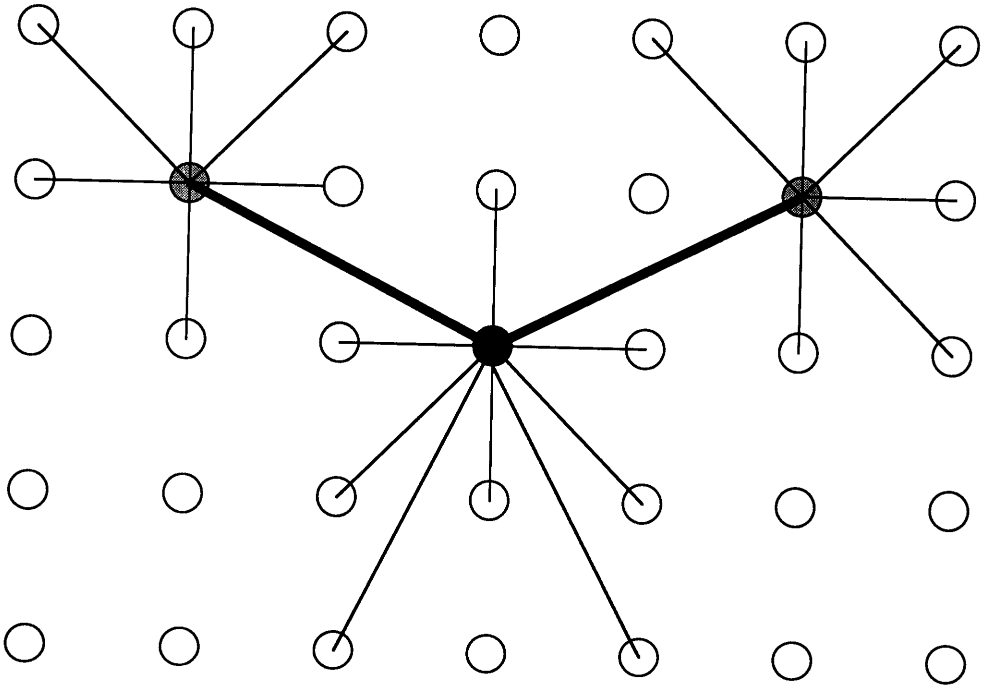


**Figure 6.** Schematic representations of the two methods used to construct the SST families. (A) The filled circle represents the family center and the lines are drawn to other members of the family that share a SST with the family center. (B) The dark-filled circle is the family center and the two grey-filled circles represent "close relatives" of the family center. Note that the family is larger because the family is the union of SSTs derived from either the family center or its close relatives.

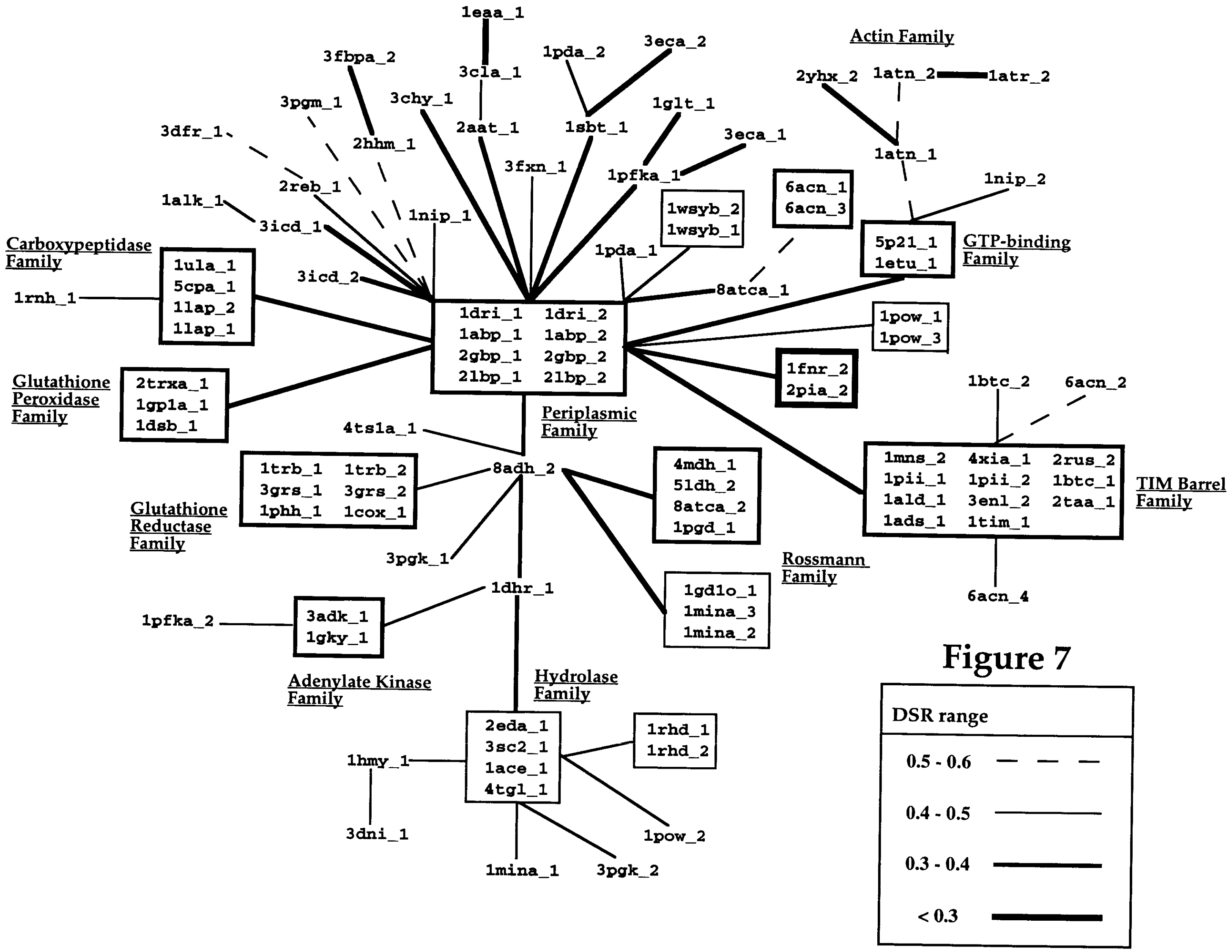

Actin Family
2yhx_2
1atn_2
1atr_2
1atn_1
1nip_2
5p21_1
1etu_1
GTP-binding Family
6acn_1
6acn_3
8atca_1
1eaa_1
3cla_1
2aat_1
1pda_2
3eca_2
1sbt_1
1glt_1
3eca_1
1pfka_1
1wsyb_2
1wsyb_1
1pda_1
3fbpa_2
2hhm_1
3pgm_1
3chy_1
3fxn_1
3dfr_1
2reb_1
1alk_1
3icd_1
1nip_1
3icd_2
Carboxypeptidase Family
1rnh_1
1ula_1
5cpa_1
1lap_2
1lap_1
1dri_1 1dri_2
1abp_1 1abp_2
2gbp_1 2gbp_2
21bp_1 21bp_2
Periplasmic Family
Glutathione Peroxidase Family
2trxa_1
1gpla_1
1dsb_1
1pow_1
1pow_3
1fnr_2
2pia_2
1btc_2
6acn_2
1mns_2 4xia_1 2rus_2
1pii_1 1pii_2 1btc_1
1ald_1 3enl_2 2taa_1
1ads_1 1tim_1
TIM Barrel Family
6acn_4
4tsla_1
8adh_2
1trb_1 1trb_2
3grs_1 3grs_2
1phh_1 1cox_1
Glutathione Reductase Family
4mdh_1
51dh_2
8atca_2
1pgd_1
Rossmann Family
1gd1o_1
1mina_3
1mina_2
3pgk_1
1dhr_1
1pfka_2
3adk_1
1gky_1
Adenylate Kinase Family
Hydrolase Family
2eda_1
3sc2_1
1ace_1
4tgl_1
1hmy_1
3dni_1
1rhd_1
1rhd_2
1pow_2
1mina_1
3pgk_2
DSR range
0.5 - 0.6
0.4 - 0.5
0.3 - 0.4
< 0.3

Figure 7

**Figure 7.** (A) Minimum spanning tree of $\alpha/\beta$ domains. The nodes are the domains and the edges represent the DSR value between the domains. The thickness of the edge indicates the strength of the similarity (see legend). Domains belonging to the same fold family have been boxed. Specific fold families have been labeled.

Figure 8A

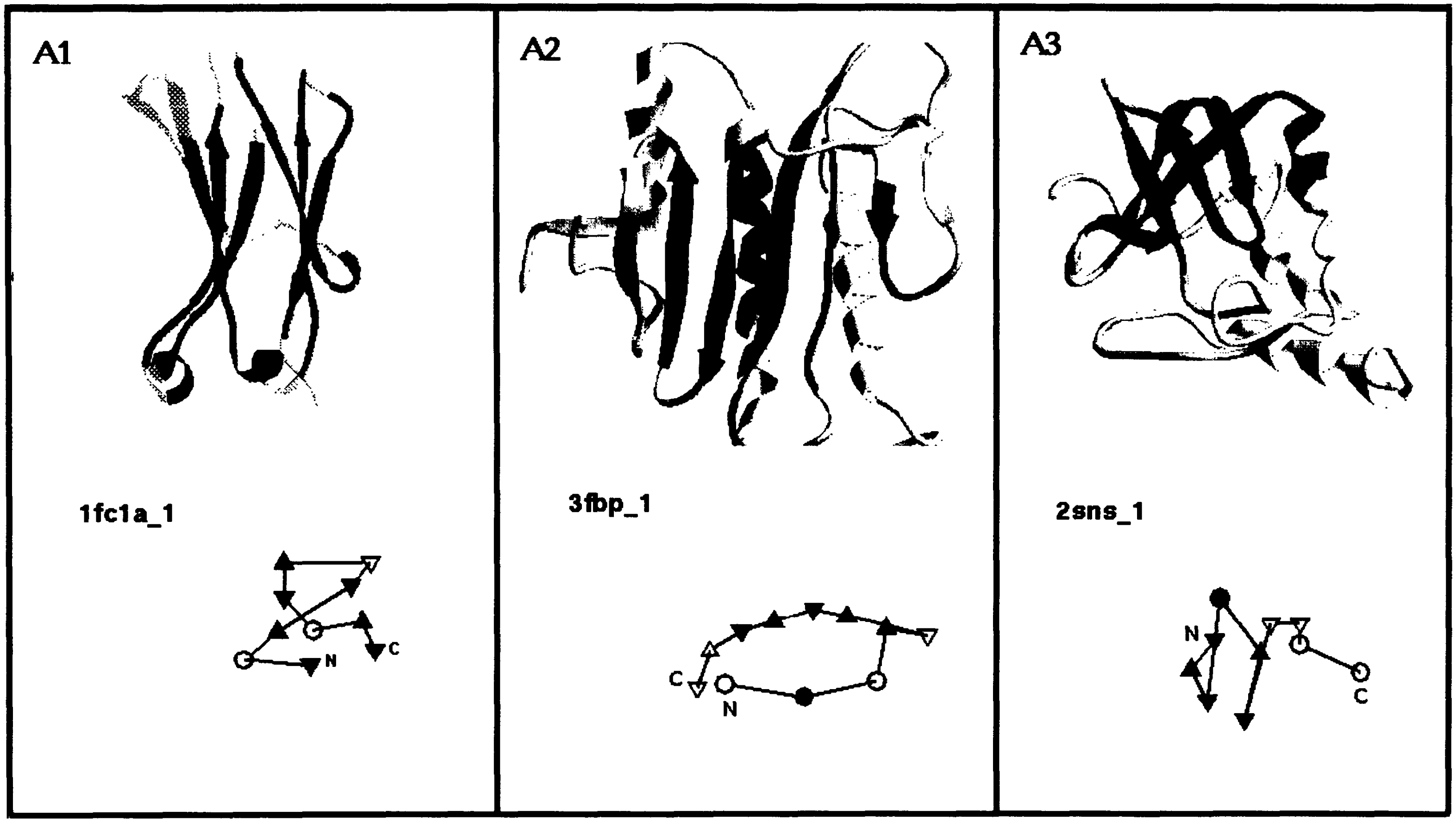


Figure 8B

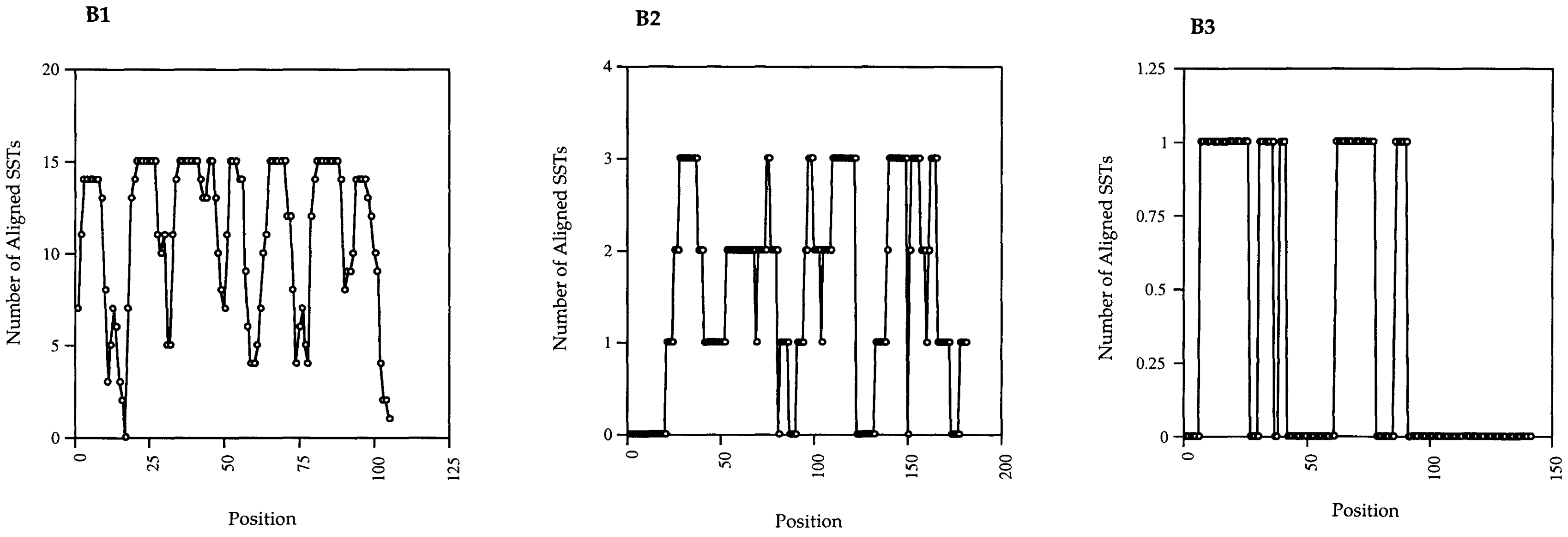
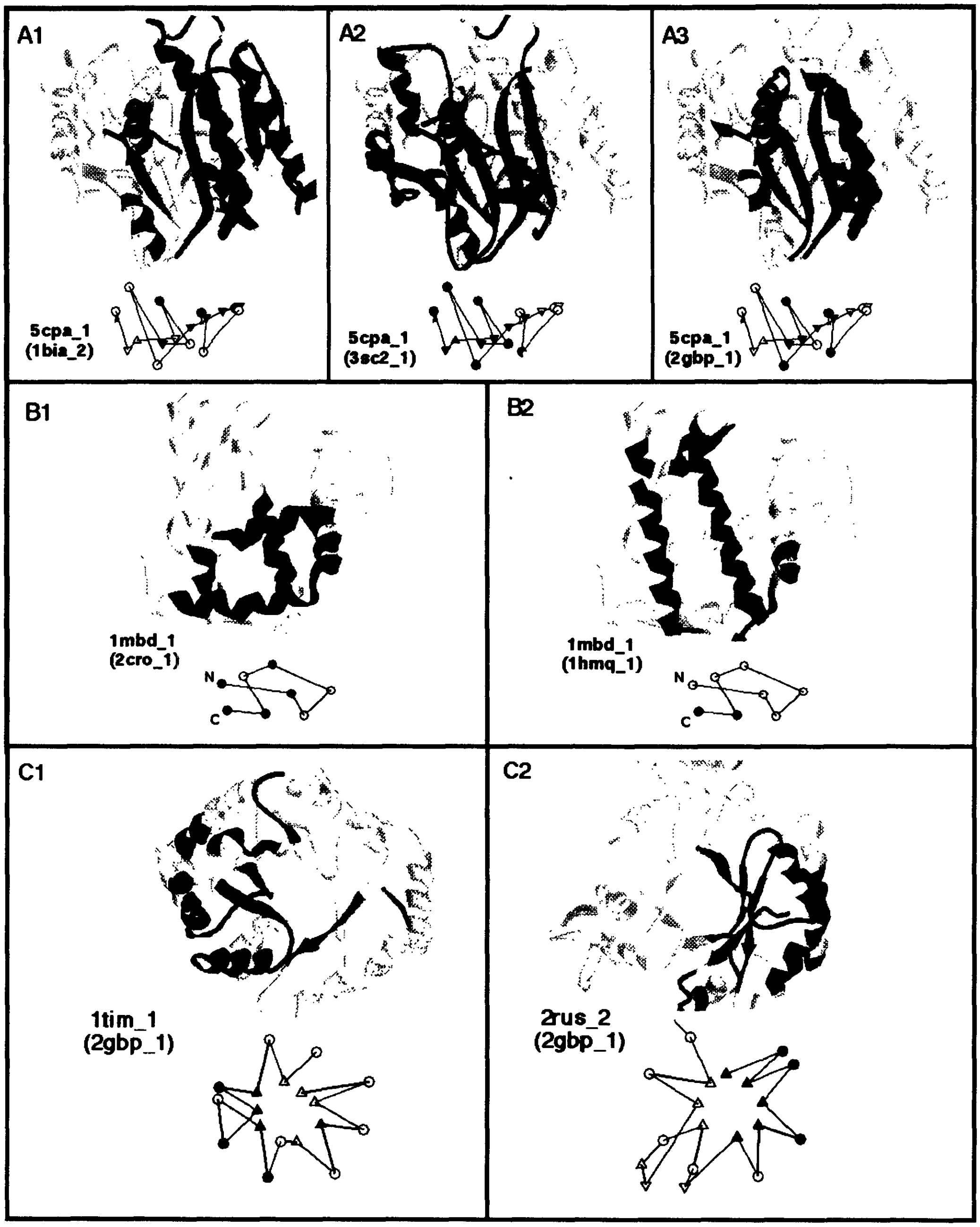

B1
20
15
10
5
0
Number of Aligned SSTs
0
25
50
75
100
125
Position
B2
4
3
2
1
0
Number of Aligned SSTs
0
50
100
150
200
Position
B3
1.25
1
0.75
0.5
0.25
0
Number of Aligned SSTs
0
50
100
150
Position

**Figure 8.** SSTs highlight the core packing region of a domain. Both (A) ribbon and topology diagrams are shown (only the core elements are shaded) along with (B) graphs of the number of aligned SSTs at each position in the structure for three domains: (1) 1fc1a_1, (2) 3fbp_1, and (3) 2sns_1.

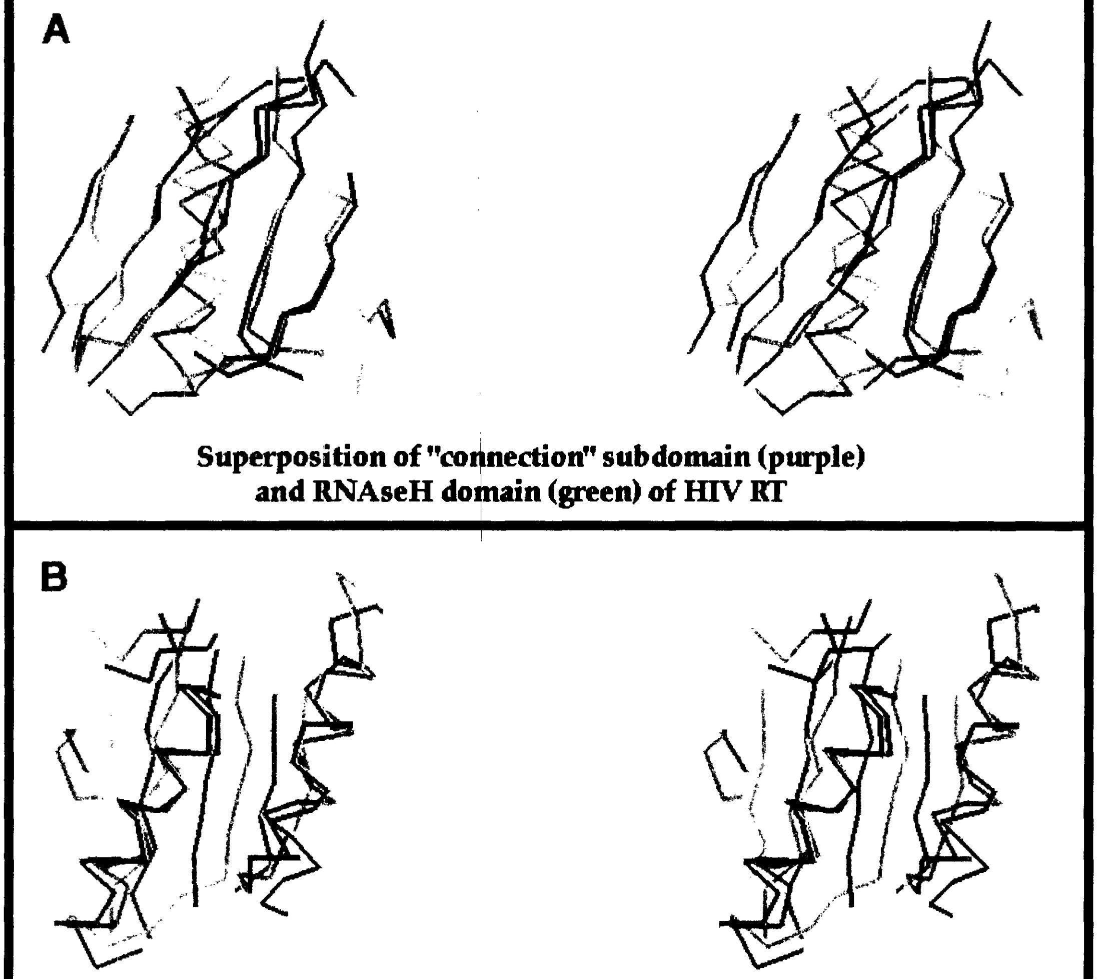
## Figure 9



A1
5cpa_1
(1bia_2)
A2
5cpa_1
(3sc2_1)
A3
5cpa_1
(2gbp_1)
B1
1mbd_1
(2cro_1)
N
C
B2
1mbd_1
(1hmq_1)
N
C
C1
1tim_1
(2gbp_1)
C2
2rus_2
(2gbp_1)

**Figure 9.** Large and complex folds can be described in terms of multiple SSTs with minimal overlap. Three examples are presented: (A) 5cpa_1 (SSTs with 3sc2_1, 1bia_2, and 2gbp_1), (B) 1mbd_1 (SSTs with 2cro_1 and 1hmq_1), and (C) 1tim_1 and 2rus_2 (SSTs with 2gbp_1). The SSTs are shaded in both the ribbon and topology diagrams.

# Figure 10



Superposition of "connection" subdomain (purple) and RNAseH domain (green) of HIV RT

Superposition of "palm" subdomain (purple) of HIV RT and 2gls_2 (green)

**Figure 10.** Identification of SSTs in HIV reverse transcriptase. (A) Alpha carbon superposition of the "connection" subdomain (purple) with the RNase H domain (green) in HIV reverse transcriptase. (B) Alpha carbon superposition of the palm subdomain (purple) with the C-terminal domain of glutamine synthetase (2gls_2, green).

Figure 11

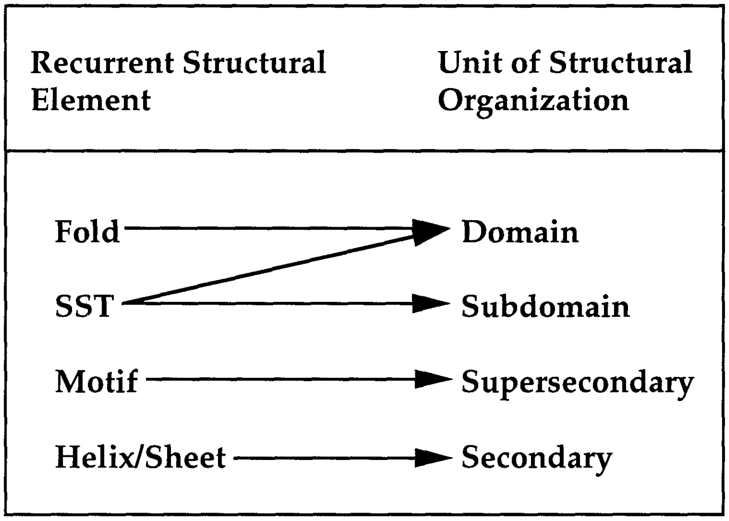


**Figure 11.** Hierarchy of protein structural organization.

# Table 1

| | Code | Title | Bounds |
|---|---|---|---|
| 1 | 155c_1 | CYTOCHROME C550 | 1-121 |
| 2 | 1aai_1 | RICIN | 1-262 |
| 3 | 1aak_1 | UBIQUITIN CONJUGATING ENZYME | 1-150 |
| 4 | 1abk_1 | ENDONUCLEASE III (1) | 1-21, 133-211 |
| 5 | 1abk_2 | ENDONUCLEASE III (2) | 22-132 |
| 6 | 1abp_1 | L-ARABINOSE-BINDING PROTEIN (1) | 1-109, 255-284 |
| 7 | 1abp_1 | L-ARABINOSE-BINDING PROTEIN (2) | 110-254, 285-306 |
| 8 | 1ace_1 | ACETYLCHOLINESTERASE | 21-480 |
| 9 | 1acx_1 | ACTINOXANTHIN | 1-108 |
| 10 | 1ads_1 | ALDOSE REDUCTASE | 1-315 |
| 11 | 1aep_1 | APOLIPOPHORIN III | 1-153 |
| 12 | 1ald_1 | ALDOLASE A | 1-363 |
| 13 | 1alk_1 | ALKALINE PHOSPHATASE (1) | 1-162, 263-449 |
| 14 | 1alk_2 | ALKALINE PHOSPHATASE (2) | 163-262 |
| 15 | 1aoz_1 | ASCORBATE OXIDASE (1) | 1-130 |
| 16 | 1aoz_2 | ASCORBATE OXIDASE (2) | 131-310 |
| 17 | 1aoz_3 | ASCORBATE OXIDASE (3) | 311-552 |
| 18 | 1arb_1 | ACHROMOBACTER PROTEASE I (1) | 1-132 |
| 19 | 1arb_2 | ACHROMOBACTER PROTEASE I (2) | 133-263 |
| 20 | 1arp_1 | PEROXIDASE (1) | 1-144, 275-291 |
| 21 | 1arp_2 | PEROXIDASE (2) | 145-274, 292-336 |
| 22 | 1atn_1 | ACTIN (1) | 1-144, 338-372 |
| 23 | 1atn_1 | ACTIN (2) | 145-337 |
| 24 | 1atr_1 | HEAT-SHOCK COGNATE PROTEIN (1) | 28-136, 230-297 |
| 25 | 1atr_2 | HEAT-SHOCK COGNATE PROTEIN (2) | 137-229, 298-383 |
| 26 | 1ayh_1 | BETA-D-GLUCAN 4-GLUCANOHYDROLASE | 1-214 |
| 27 | 1baa_1 | ENDOCHITINASE | 1-243 |
| 28 | 1ban_1 | BARNASE | 1-108 |
| 29 | 1bbp_1 | BILIN BINDING PROTEIN | 1-173 |
| 30 | 1bgc_1 | GRANULOCYTE CSF | 1-158 |
| 31 | 1bia_1 | BIRA BIFUNCTIONAL PROTEIN (1) | 1-60 |
| 32 | 1bia_2 | BIRA BIFUNCTIONAL PROTEIN (2) | 61-247 |
| 33 | 1bmv2_1 | BEAN POD MOTTLE VIRUS (1) | 1-182 |
| 34 | 1bmv2_2 | BEAN POD MOTTLE VIRUS (2) | 183-374 |
| 35 | 1bov_1 | VEROTOXIN-1 | 1-69 |
| 36 | 1bp2_1 | PHOSPHOLIPASE A2 | 1-123 |
| 37 | 1btc_1 | BETA-AMYLASE (1) | 1-266 |
| 38 | 1btc_2 | BETA-AMYLASE (2) | 267-491 |
| 39 | 1cc5_1 | CYTOCHROME C5 | 1-83 |
| 40 | 1cd8_1 | CD8 | 1-114 |
| 41 | 1cew_1 | CYSTATIN | 1-108 |
| 42 | 1cid_1 | CD4 DOMAINS 3 AND 4 (1) | 1-106 |
| 43 | 1cid_2 | CD4 DOMAINS 3 AND 4 (2) | 107-177 |
| 44 | 1cmb_1 | MET APO-REPRESSOR | 1-104 |
| 45 | 1col_1 | COLICIN A | 1-197 |
| 46 | 1cox_1 | CHOLESTEROL OXIDASE | 1-115, 191-280 |
| 47 | 1cox_2 | CHOLESTEROL OXIDASE | 116-190, 281-502 |
| 48 | 1cpca_1 | C-PHYCOCYANIN (ALPHA) | 1-162 |
| 49 | 1cpcb_1 | C-PHYCOCYANIN (BETA) | 1-172 |
| 50 | 1ctf_1 | L7/L12 50S RIBOSOMAL PROTEIN | 1-78 |
| 51 | 1dhr_1 | DIHYDROPTERIDINE REDUCTASE | 1-236 |
| 52 | 1dri_1 | D-RIBOSE-BINDING PROTEIN | 1-103, 235-263 |
| 53 | 1dri_2 | D-RIBOSE-BINDING PROTEIN | 104-234, 264-271 |
| 54 | 1dsb_1 | DISULFIDE BOND FORM. PROT. (1) | 1-62, 139-188 |
| 55 | 1dsb_2 | DISULFIDE BOND FORM. PROT. (2) | 63-138 |
| 56 | 1eaa_1 | DIHYDROLIPOYL TRANSACETYLASE | 1-243 |
| 57 | 1ecd_1 | HEMOGLOBIN (ERYTHROCRUORIN) | 1-136 |
| 58 | 1end_1 | ENDONUCLEASE V | 1-137 |
| 59 | 1etu_1 | ELONGATION FACTOR TU | 1-141 |
| 60 | 1f3g_1 | PHOSPHOCARRIER III | 1-150 |
| 61 | 1fc1a_1 | FC FRAGMENT IGG1 CLASS (1) | 1-105 |
| 62 | 1fc1a_2 | FC FRAGMENT IGG1 CLASS (2) | 106-206 |
| 63 | 1fha_1 | FERRITIN (H-CHAIN) | 1-170 |
| 64 | 1fkf_1 | FK506 BINDING PROTEIN | 1-107 |
| 65 | 1fnr_1 | FERREDOXIN REDUCTASE (1) | 1-141 |
| 66 | 1fnr_2 | FERREDOXIN REDUCTASE (2) | 142-296 |
| 67 | 1gcr_1 | GAMMA IVA-CRYSTALLIN (1) | 1-85 |
| 68 | 1gcr_2 | GAMMA IVA-CRYSTALLIN (2) | 86-174 |
| 69 | 1gd1o_1 | HOLO-D-G3P DEHYDROGENASE (1) | 1-147 |
| 70 | 1gd1o_2 | HOLO-D-G3P DEHYDROGENASE (2) | 148-334 |
| 71 | 1gky_1 | GUANYLATE KINASE | 1-186 |
| 72 | 1glt_1 | GLUTATHIONE SYNTHASE | 1-121 |
| 73 | 1gly_1 | GLUCOAMYLASE (1) | 1-226 |
| 74 | 1gly_2 | GLUCOAMYLASE (2) | 227-470 |
| 75 | 1gmf_1 | GRANULOCYTE-MACROPHAGE CSF | 1-119 |
| 76 | 1gmp_1 | RNASE SA | 1-96 |
| 77 | 1gox_1 | GLYCOLATE OXIDASE | 1-350 |
| 78 | 1gp1_1 | GLUTATHIONE PEROXIDASE | 1-184 |
| 79 | 1hip_1 | HIGH POTENTIAL IRON PROTEIN | 1-85 |
| 80 | 1hmq_1 | HEMERYTHRIN | 1-113 |
| 81 | 1hmy_1 | HHAI METHYLTRANSFERASE (1) | 1-193, 304-327 |
| 82 | 1hmy_2 | HHAI METHYLTRANSFERASE (2) | 194-280 |
| 83 | 1hoe_1 | ALPHA-AMYLASE INHIBITOR HOE-467 | 1-74 |
| 84 | 1hst_1 | HISTONE H5 | 1-74 |
| 85 | 1huw_1 | HUMAN GROWTH HORMONE | 1-166 |
| 86 | 1i1b_1 | INTERLEUKIN-1 BETA | 1-151 |
| 87 | 1ifb_1 | FATTY ACID BINDING PROTEIN | 1-131 |
| 88 | 1lap_1 | LEUCINE AMINOPEPTIDASE (1) | 1-155 |
| 89 | 1lap_2 | LEUCINE AMINOPEPTIDASE (2) | 156-481 |
| 90 | 1lfb_1 | TRANS. FACTOR LFB1 (HOMEO) | 1-78 |
| 91 | 1lh1_1 | LEGHEMOGLOBIN | 1-153 |
| 92 | 1lmb_1 | LAMBDA REPRESSOR | 1-92 |
| 93 | 1lpe_1 | APOLIPOPROTEIN-E3 | 1-144 |
| 94 | 1lts_1 | HEAT-LABILE ENTEROTOXIN | 1-103 |
| 95 | 1lz1_1 | LYSOZYME | 1-130 |
| 96 | 1mat_1 | METHIONINE AMINOPEPTIDASE (1) | 1-119 |
| 97 | 1mat_2 | METHIONINE AMINOPEPTIDASE (2) | 120-263 |
| 98 | 1mbd_1 | MYOGLOBIN | 1-153 |
| 99 | 1mcpl_1 | IMMUNOGLOBULIN FAB FRAGMENT (1) | 1-110 |
| 100 | 1mcpl_2 | IMMUNOGLOBULIN FAB FRAGMENT (2) | 111-220 |
| 101 | 1min_1 | NITROGENASE MO-FE PROTEIN (1) | 50-202 |
| 102 | 1min_2 | NITROGENASE MO-FE PROTEIN (2) | 203-318 |
| 103 | 1min_3 | NITROGENASE MO-FE PROTEIN (3) | 1-49, 319-470 |
| 104 | 1mns_1 | MANDELATE RACEMASE (1) | 1-130 |
| 105 | 1mns_2 | MANDELATE RACEMASE (2) | 131-357 |
| 106 | 1ms2_1 | MS2 VIRUS COAT PROTEIN | 1-129 |
| 107 | 1msb_1 | MANNOSE BINDING PROTEIN | 1-115 |
| 108 | 1mup_1 | MAJOR URINARY PROTEIN | 1-157 |
| 109 | 1mvp_1 | MYELOBLASTOSIS VIRAL PROTEASE | 1-112 |
| 110 | 1ndk_1 | NUCLEOSIDE DIPHOSPHATE KINASE | 1-148 |
| 111 | 1nip_1 | NITROGENASE IRON PROTEIN (1) | 1-146 |
| 112 | 1nip_2 | NITROGENASE IRON PROTEIN (2) | 147-283 |
| 113 | 1nsb_1 | NEURAMINIDASE (SIALIDASE) | 1-390 |
| 114 | 1onc_1 | P-30 PROTEIN | 1-103 |
| 115 | 1ova_1 | OVALBUMIN | 1-385 |
| 116 | 1paz_1 | PSEUDOAZURIN | 1-120 |
| 117 | 1pcy_1 | PLASTOCYANIN | 1-99 |
| 118 | 1pda_1 | PORPHOBILINOGEN DEAMINASE (1) | 1-91, 189-206 |
| 119 | 1pda_2 | PORPHOBILINOGEN DEAMINASE (2) | 92-188 |
| 120 | 1pda_3 | PORPHOBILINOGEN DEAMINASE (3) | 207-296 |
| 121 | 1pfk_1 | PHOSPHOFRUCTOKINASE (1) | 1-130, 252-305 |
| 122 | 1pfk_2 | PHOSPHOFRUCTOKINASE (2) | 131-251, 306-320 |
| 123 | 1pgd_1 | 6-PHOSPHOGLUCONATE DEHYDR. (1) | 1-176 |
| 124 | 1pgd_2 | 6-PHOSPHOGLUCONATE DEHYDR. (2) | 177-469 |
| 125 | 1pgx_1 | PROTEIN G TYPE 7 | 1-70 |
| 126 | 1phh_1 | P-HYDROXYBENZOATE HYDROXY. (1) | 1-175 |
| 127 | 1phh_2 | P-HYDROXYBENZOATE HYDROXY. (2) | 176-290 |
| 128 | 1phh_3 | P-HYDROXYBENZOATE HYDROXY. (3) | 291-394 |
| 129 | 1pii_1 | ANTHRANILATE ISOMERASE | 1-255 |
| 130 | 1pii_2 | INDOLE-3-GLYCEROL-P SYNTHASE | 256-452 |
| 131 | 1pkp_1 | RIBOSOMAL PROTEIN S5 | 1-145 |
| 132 | 1plf_1 | PLATELET FACTOR 4 | 1-65 |
| 133 | 1poh_1 | HIS PHOSPHOCARRIER PROTEIN | 1-85 |
| 134 | 1pow_1 | PYRUVATE OXIDASE (1) | 1-191 |
| 135 | 1pow_2 | PYRUVATE OXIDASE (2) | 192-342 |
| 136 | 1pow_3 | PYRUVATE OXIDASE (3) | 343-593 |
| 137 | 1pts_1 | STREPTAVIDIN | 1-119 |
| 138 | 1pyp_1 | INORGANIC PYROPHOSPHATASE | 1-280 |
| 139 | 1rbp_1 | RETINOL BINDING PROTEIN | 1-174 |
| 140 | 1rec_1 | RECOVERIN (1) | 1-84 |
| 141 | 1rec_2 | RECOVERIN (2) | 85-185 |
| 142 | 1rhd_1 | RHODANESE (1) | 1-150 |
| 143 | 1rhd_2 | RHODANESE (2) | 151-293 |
| 144 | 1rib_1 | R2 OF RIBONUCLEOTIDE REDUCTASE | 1-340 |
| 145 | 1rn3_1 | RIBONUCLEASE A | 1-124 |
| 146 | 1rnh_1 | SELENOMET. RIBONUCLEASE H | 1-148 |
| 147 | 1rnt_1 | RIBONUCLEASE T1 | 1-104 |
| 148 | 1sbt_1 | SUBTILISIN BPN | 1-275 |
| 149 | 1sgt_1 | TRYPSIN SGT (1) | 1-103 |
| 150 | 1sgt_2 | TRYPSIN SGT (2) | 104-223 |
| 151 | 1sha_1 | V-SRC SH2 DOMAIN | 1-59 |
| 152 | 1shf_1 | FYN PROTO-ONCOGENE SH3 DOMAIN | 1-59 |
| 153 | 1sil_1 | SIALIDASE | 1-381 |
| 154 | 1sn3_1 | SCORPION NEUROTOXIN | 1-65 |
| 155 | 1sry_1 | SERYL-TRNA SYNTHETASE | 101-421 |
| 156 | 1tbp_1 | TATA-BINDING PROTEIN (1) | 1-90 |
| 157 | 1tbp_2 | TATA-BINDING PROTEIN (2) | 91-180 |
| 158 | 1ten_1 | TENASCIN | 1-89 |
| 159 | 1tfg_1 | TGF TYPE BETA 2 | 1-112 |
| 160 | 1tie_1 | KUNITZ TRYPSIN INHIBITOR | 1-167 |
| 161 | 1tim_1 | TRIOSE PHOSPHATE ISOMERASE | 1-247 |
| 162 | 1tlk_1 | TELOKIN | 1-103 |
| 163 | 1tnf_1 | TUMOR NECROSIS FACTOR-ALPHA | 1-152 |
| 164 | 1trb_1 | THIOREDOXIN REDUCTASE | 1-114 |
| 165 | 1trb_2 | THIOREDOXIN REDUCTASE | 115-244 |
| 166 | 1ubq_1 | UBIQUITIN | 1-76 |
| 167 | 1ula_1 | PURINE NUCLEO. PHOSPHORYLASE | 1-289 |
| 168 | 1vsg_1 | VARIANT SURFACE GLYCOPROTEIN | 1-362 |
| 169 | 1wsyb_1 | TRYPTOPHAN SYNTHASE | 1-44, 78-196 |
| 170 | 1wsyb_2 | TRYPTOPHAN SYNTHASE | 45-77, 197-385 |
| 171 | 1zaa_1 | ZIF268 IMMEDIATE EARLY GENE | 1-85 |
| 172 | 256b_1 | CYTOCHROME B562 | 1-106 |
| 173 | 2aat_1 | ASPARTATE AMINOTRANSFERASE | 1-315 |
| 174 | 2aat_2 | ASPARTATE AMINOTRANSFERASE | 316-396 |
| 175 | 2alp_1 | ALPHA-LYTIC PROTEASE (1) | 1-87 |
| 176 | 2alp_2 | ALPHA-LYTIC PROTEASE (2) | 88-198 |
| 177 | 2apr_1 | RHIZOPUSPEPSIN (1) | 1-173 |
| 178 | 2apr_2 | RHIZOPUSPEPSIN (2) | 174-325 |
| 179 | 2aza_1 | AZURIN | 1-129 |
| 180 | 2bbk_1 | METHYLAMINE DEHYDROGENASE | 1-355 |
| 181 | 2bop_1 | BPV E2 DNA-BINDING DOMAIN | 1-85 |
| 182 | 2bpa1_1 | PHIX174 CAPSID PROTEINS | 1-426 |
| 183 | 2cab_1 | CARBONIC ANHYDRASE FORM B | 1-256 |
| 184 | 2ccy_1 | CYTOCHROME C' | 1-127 |
| 185 | 2cd4_1 | CD4 N-TERMINAL FRAGMENT (1) | 1-97 |
| 186 | 2cd4_2 | CD4 N-TERMINAL FRAGMENT (2) | 98-176 |
| 187 | 2ci2i_1 | CHYMOTRYPSIN INHIBITOR 2 | 1-75 |
| 188 | 2cpk_1 | CAMP PROTEIN KINASE (1) | 1-125, 303-336 |
| 189 | 2cpk_2 | CAMP PROTEIN KINASE (2) | 126-302 |
| 190 | 2cpl_1 | CYCLOPHILIN A | 1-164 |
| 191 | 2cpp_1 | CYTOCHROME P450CAM (1) | 1-95, 295-340 |
| 192 | 2cpp_2 | CYTOCHROME P450CAM (2) | 96-294, 341-405 |
| 193 | 2cro_1 | 434 CRO PROTEIN | 1-65 |
| 194 | 2cts_1 | CITRATE SYNTHASE (1) | 1-274, 381-437 |
| 195 | 2cts_2 | CITRATE SYNTHASE (2) | 275-380 |
| 196 | 2cyp_1 | CYTOCHROME C PEROXIDASE (1) | 1-160 |
| 197 | 2cyp_2 | CYTOCHROME C PEROXIDASE (2) | 161-345 |
| 198 | 2eda_1 | HALOALKANE DEHALOGENASE (1) | 1-155, 230-310 |
| 199 | 2eda_2 | HALOALKANE DEHALOGENASE (2) | 156-229 |
| 200 | 2fgf_1 | BASIC FIBROBLAST GROWTH FACTOR | 1-126 |
| 201 | 2gbp_1 | D-GAL-GLU BINDING PROTEIN (1) | 1-110, 257-294 |
| 202 | 2gbp_2 | D-GAL-GLU BINDING PROTEIN (2) | 111-256, 295-309 |
| 203 | 2gls_1 | GLUTAMINE SYNTHETASE (1) | 1-103 |
| 204 | 2gls_2 | GLUTAMINE SYNTHETASE (2) | 104-468 |
| 205 | 2gst_1 | GLUTATHIONE S-TRANSFERASE (1) | 1-89 |
| 206 | 2gst_2 | GLUTATHIONE S-TRANSFERASE (2) | 90-217 |
| 207 | 2hhbb_1 | HEMOGLOBIN BETA | 1-146 |
| 208 | 2hhm_1 | HUMAN INOSITOL MONOPHOSPHATASE | 1-272 |
| 209 | 2hlaa_1 | CLASS I HISTOCOMPATIBILITY (1) | 1-181 |
| 210 | 2hlaa_2 | CLASS I HISTOCOMPATIBILITY (2) | 182-270 |
| 211 | 2hlab_1 | BETA-MICROGLOBULIN | 1-99 |
| 212 | 2hmga_1 | HEMAGGLUTININ | 1-175 |
| 213 | 2hpd_1 | CYTOCHROME P450 (1) | 72-325, 390-457 |
| 214 | 2hpd_2 | CYTOCHROME P450 (2) | 1-71, 326-389 |
| 215 | 2lbp_1 | LEUCINE-BINDING PROTEIN (1) | 1-118, 253-326 |
| 216 | 2lbp_2 | LEUCINE-BINDING PROTEIN (2) | 124-247, 332-344 |
| 217 | 2lzm_1 | LYSOZYME (1) | 1-73 |
| 218 | 2lzm_2 | LYSOZYME (2) | 74-164 |
| 219 | 2mev1_1 | MENGO VIRUS COAT PROTEIN | 1-268 |
| 220 | 2pab_1 | PREALBUMIN | 1-124 |
| 221 | 2pia_1 | PHTHALATE DIOXYGENASE (1) | 1-111 |
| 222 | 2pia_2 | PHTHALATE DIOXYGENASE (2) | 112-226 |
| 223 | 2pia_3 | PHTHALATE DIOXYGENASE (3) | 227-321 |
| 224 | 2reb_1 | RECA PROTEIN | 1-303 |
| 225 | 2rus_1 | RUBISCO (1) | 1-148 |
| 226 | 2rus_2 | RUBISCO (2) | 149-437 |
| 227 | 2rve_1 | ECORV ENDONUCLEASE | 1-212 |
| 228 | 2scp_1 | SARCOPLASM CA-BIND. PROT. (1) | 1-84 |
| 229 | 2scp_2 | SARCOPLASM CA-BIND. PROT. (2) | 85-174 |
| 230 | 2sns_1 | STAPHYLOCOCCAL NUCLEASE | 1-141 |
| 231 | 2snv_1 | SINDBIS VIRUS CAPSID PROT. (1) | 1-66 |
| 232 | 2snv_2 | SINDBIS VIRUS CAPSID PROT. (2) | 67-151 |
| 233 | 2sod_1 | CU,ZN SUPEROXIDE DISMUTASE | 1-151 |
| 234 | 2ssi_1 | STREP. SUBTILISIN INHIBITOR | 1-107 |
| 235 | 2stv_1 | TOBACCO NECROSIS VIRUS | 1-184 |
| 236 | 2taa_1 | TAKA-AMYLASE A | 1-360 |
| 237 | 2tbvc_1 | TOMATO BUSHY STUNT VIRUS (1) | 1-211 |
| 238 | 2tbvc_2 | TOMATO BUSHY STUNT VIRUS (2) | 212-321 |
| 239 | 2tmv_1 | TOBACCO MOSAIC VIRUS | 1-154 |
| 240 | 2trx_1 | THIOREDOXIN | 1-108 |
| 241 | 2yhx_1 | YEAST HEXOKINASE B (1) | 1-50, 191-430 |
| 242 | 2yhx_2 | YEAST HEXOKINASE B (2) | 51-190, 431-457 |
| 243 | 351c_1 | CYTOCHROME C551 | 1-82 |
| 244 | 3adk_1 | ADENYLATE KINASE | 1-194 |
| 245 | 3b5c_1 | CYTOCHROME B5 | 1-85 |
| 246 | 3bcl_1 | BACTERIOCHLOROPHYLL-A PROTEIN | 1-344 |
| 247 | 3blm_1 | BETA-LACTAMASE | 1-30, 154-257 |
| 248 | 3blm_1 | BETA-LACTAMASE | 31-153 |
| 249 | 3chy_1 | CHEY | 1-128 |
| 250 | 3cla_1 | CHLORAMPHEN. ACETYLTRANSFER. | 1-213 |
| 251 | 3cln_1 | CALMODULIN (1) | 1-75 |
| 252 | 3cln_2 | CALMODULIN (2) | 76-143 |
| 253 | 3cna_1 | CONCANAVALIN A | 1-237 |
| 254 | 3cpv_1 | CALCIUM-BINDING PARVALBUMIN | 1-108 |
| 255 | 3dfr_1 | DIHYDROFOLATE REDUCTASE | 1-162 |
| 256 | 3dni_1 | DEOXYRIBONUCLEASE I | 1-259 |
| 257 | 3eca_1 | ASPARAGINASE TYPE II (1) | 1-200 |
| 258 | 3eca_2 | ASPARAGINASE TYPE II (2) | 201-326 |
| 259 | 3enl_1 | ENOLASE (1) | 1-142 |
| 260 | 3enl_2 | ENOLASE (2) | 143-436 |
| 261 | 3fbp_1 | FRUCTOSE-1,6-BISPHOSPHATASE (1) | 1-181 |
| 262 | 3fbp_2 | FRUCTOSE-1,6-BISPHOSPHATASE (2) | 182-316 |
| 263 | 3fxc_1 | FERREDOXIN | 1-98 |
| 264 | 3fxn_1 | FLAVODOXIN | 1-138 |
| 265 | 3gap_1 | CATABOLITE ACTIVATOR PROT. (1) | 1-135 |
| 266 | 3gap_2 | CATABOLITE ACTIVATOR PROT. (2) | 136-209 |
| 267 | 3grs_1 | GLUTATHIONE REDUCTASE (1) | 1-139 |
| 268 | 3grs_2 | GLUTATHIONE REDUCTASE (2) | 140-272 |
| 269 | 3grs_3 | GLUTATHIONE REDUCTASE (3) | 345-461 |
| 270 | 3hmg_1 | HEMAGGLUTININ | 1-328 |
| 271 | 3hvp_1 | HIV-1 PROTEASE | 1-99 |
| 272 | 3icb_1 | CALCIUM-BINDING PROTEIN | 1-75 |
| 273 | 3icd_1 | ISOCITRATE DEHYDROGENASE (1) | 1-123, 318-414 |
| 274 | 3icd_2 | ISOCITRATE DEHYDROGENASE (2) | 124-157, 203-317 |
| 275 | 3il8_1 | INTERLEUKIN 8 | 1-68 |
| 276 | 3ink_1 | INTERLEUKIN 2 | 1-122 |
| 277 | 3mds_1 | MAN. SUPEROXIDE DISMUTASE (1) | 1-94 |
| 278 | 3mds_2 | MAN. SUPEROXIDE DISMUTASE (2) | 95-203 |
| 279 | 3pgk_1 | PHOSPHOGLYCERATE KINASE (1) | 1-192 |
| 280 | 3pgk_2 | PHOSPHOGLYCERATE KINASE (2) | 193-415 |
| 281 | 3pgm_1 | PHOSPHOGLYCERATE MUTASE | 1-230 |
| 282 | 3rp2_1 | RAT MAST CELL PROTEASE II (1) | 1-114 |
| 283 | 3rp2_2 | RAT MAST CELL PROTEASE II (2) | 115-224 |
| 284 | 3sc2_1 | SERINE CARBOXYPEPTIDASE II | 1-254 |
| 285 | 3tln_1 | THERMOLYSIN (1) | 1-151 |
| 286 | 3tln_2 | THERMOLYSIN (2) | 152-316 |
| 287 | 3wrp_1 | TRP APOREPRESSOR | 1-101 |
| 288 | 4fd1_1 | FERREDOXIN | 1-106 |
| 289 | 4fis_1 | FIS PROTEIN | 1-73 |
| 290 | 4mdh_1 | CYT. MALATE DEHYDROGENASE (1) | 1-151 |
| 291 | 4mdh_2 | CYT. MALATE DEHYDROGENASE (2) | 152-333 |
| 292 | 4sbv_1 | SOUTH. BEAN MOSAIC VIRUS COAT | 1-222 |
| 293 | 4tgl_2 | TRIACYLGLYCEROL ACYLHYDROLASE | 1-265 |
| 294 | 4tms_1 | THYMIDYLATE SYNTHASE (1) | 70-140 |
| 295 | 4tms_2 | THYMIDYLATE SYNTHASE (2) | 1-69, 141-316 |
| 296 | 4ts1_1 | TYROSYL-TRNA SYNTHETASE (1) | 1-230 |
| 297 | 4ts1_2 | TYROSYL-TRNA SYNTHETASE (2) | 231-317 |
| 298 | 4xia_1 | D-XYLOSE ISOMERASE | 1-393 |
| 299 | 5cpa_1 | CARBOXYPEPTIDASE A | 1-307 |
| 300 | 5ldh_1 | LACTATE DEHYDROGENASE (1) | 1-21, 165-331 |
| 301 | 5ldh_2 | LACTATE DEHYDROGENASE (2) | 22-164 |
| 302 | 5p21_1 | C-H-RAS P21 PROTEIN | 1-166 |
| 303 | 6acn_1 | ACONITASE (1) | 1-201 |
| 304 | 6acn_2 | ACONITASE (2) | 202-319 |
| 305 | 6acn_3 | ACONITASE (3) | 320-512 |
| 306 | 6acn_4 | ACONITASE (4) | 537-753 |
| 307 | 7api_1 | ALPHA-1-ANTITRYPSIN | 1-339 |
| 308 | 7wga_1 | WHEAT GERM AGGLUTININ | 1-171 |
| 309 | 8adh_1 | ALCOHOL DEHYDROGENASE (1) | 1-174, 319-374 |
| 310 | 8adh_2 | ALCOHOL DEHYDROGENASE (2) 1 | 75-318 |
| 311 | 8atca_1 | ASPARTATE CARBAMOYLTRANS. (1) | 1-151 |
| 312 | 8atca_2 | ASPARTATE CARBAMOYLTRANS. (2) | 152-310 |
| 313 | 8atcb_1 | ASPARTATE CARBAMOYL. (REG.) | 1-146 |
| 314 | 8cat_1 | CATALASE (1) | 74-318 |
| 315 | 8cat_2 | CATALASE (2) | 319-434 |
| 316 | 9pap_1 | PAPAIN (1) | 19-112, 208-212 |
| 317 | 9pap_2 | PAPAIN (2) | 1-18, 113-207 |

**Table 1.** List of domains used in this study. The PDB code, title, and residue boundaries of each domain are listed.

# Table 2

**Relationship between minimum alignment length and the number of significant similarities**

| Minimum Alignment Length | Number of Comparisons | |
|---|---|---|
| | nDRMS ≤ 0.6 | nDRMS ≤ 0.5 |
| 0.35 | 2978 | 1391 |
| 0.40 | 2511 | 1141 |
| 0.45 | 2035 | 957 |
| 0.50 | 1548 | 752 |
| 0.55 | 1145 | 609 |
| 0.60 | 778 | 459 |
| 0.65 | 532 | 335 |
| Lmin | 1926 | 899 |

**Table 2.** Relationship between minimum alignment length and the number of significant comparisons. The alignment length L is reported as a fraction of the length of the smaller domain, $n_S$. Lmin refers to the minimum alignment length function described in the text. The two columns present the number of comparisons with a nDRMS($Y_L$) score less than or equal to 0.5 or 0.6, given the specified minimum alignment length.

# Table 3

| | | Multiple-center | | Single-center | |
|---|---|---|---|---|---|
| | Domain | DSR | Closest relative | DSR (2gbp_1) | Structural Family |
| 1 | 2gbp_1 | 0.00 | 2gbp_1 | 0.00 | Periplasmic |
| 2 | 1dri_1 | 0.00 | 1dri_1 | 0.13 | Periplasmic |
| 3 | 1abp_1 | 0.00 | 1abp_1 | 0.23 | Periplasmic |
| 4 | 21bp_1 | 0.00 | 21bp_1 | 0.28 | Periplasmic |
| 5 | 1dri_2 | 0.00 | 1dri_2 | 0.35 | Periplasmic |
| 6 | 2gbp_2 | 0.00 | 2gbp_2 | 0.40 | Periplasmic |
| 7 | 21bp_2 | 0.00 | 21bp_2 | 0.47 | Periplasmic |
| 8 | 3chy_1 | 0.00 | 3chy_1 | 0.48 | |
| 9 | 1abp_2 | 0.25 | 1dri_2 | 0.51 | Periplasmic |
| 10 | 1pfk_1 | 0.25 | 1dri_1 | 0.37 | |
| 11 | 5p21_1 | 0.31 | 1dri_1 | 0.31 | Ras/GTP-binding |
| 12 | 8atca_1 | 0.32 | 1dri_1 | 0.41 | Rossmann fold |
| 13 | 1sbt_1 | 0.32 | 2gbp_1 | 0.32 | Subtilisin |
| 14 | 8adh_2 | 0.34 | 1dri_2 | 0.56 | Rossmann fold |
| 15 | 3icd_1 | 0.35 | 21bp_1 | 0.44 | |
| 16 | 1pgd_1 | 0.35 | 3chy_1 | 0.41 | Rossmann fold |
| 17 | 3icd_2 | 0.37 | 2gbp_2 | 0.65 | |
| 18 | 2pia_2 | 0.37 | 21bp_2 | 0.49 | |
| 19 | 1min_3 | 0.38 | 3chy_1 | 0.57 | |
| 20 | 1gd1o_1 | 0.38 | 21bp_2 | 0.41 | Rossmann fold |
| 21 | 3eca_1 | 0.39 | 1dri_1 | 0.45 | |
| 22 | 8atca_2 | 0.39 | 2gbp_1 | 0.39 | Rossmann fold |
| 23 | 1ula_1 | 0.39 | 1dri_1 | 0.44 | Carboxypeptidase |
| 24 | 3eca_2 | 0.39 | 1dri_1 | 0.41 | |
| 25 | 1mns_2 | 0.40 | 1dri_1 | 0.44 | TIM barrel |
| 26 | 3fxn_1 | 0.40 | 21bp_2 | 0.47 | |
| 27 | 1pow_3 | 0.41 | 21bp_2 | 0.58 | |
| 28 | 1dhr_1 | 0.42 | 1dri_2 | 0.44 | |
| 29 | 2aat_1 | 0.42 | 2gbp_1 | 0.42 | |
| 30 | 1wsyb_2 | 0.42 | 1dri_2 | 0.54 | |
| 31 | 2reb_1 | 0.43 | 1dri_2 | 0.71 | |
| 32 | 1glt_1 | 0.44 | 1dri_1 | 0.47 | |
| 33 | 1ace_1 | 0.44 | 21bp_2 | 0.49 | Hydrolase |
| 34 | 51dh_2 | 0.44 | 21bp_2 | 0.58 | Rossmann fold |
| 35 | 4ts1_1 | 0.45 | 2gbp_2 | 0.55 | Rossmann fold |
| 36 | 1min_2 | 0.45 | 3chy_1 | 0.54 | |
| 37 | 2eda_1 | 0.45 | 1dri_1 | 0.48 | |
| 38 | 2rus_2 | 0.46 | 2gbp_1 | 0.46 | TIM barrel |
| 39 | 4xia_1 | 0.46 | 2gbp_1 | 0.46 | TIM barrel |
| 40 | 1fnr_2 | 0.46 | 21bp_2 | 0.53 | |
| 41 | 1lap_2 | 0.46 | 2gbp_1 | 0.46 | Carboxypeptidase |
| 42 | 1lap_1 | 0.46 | 2gbp_1 | 0.46 | Carboxypeptidase |
| 43 | 5cpa_1 | 0.46 | 2gbp_1 | 0.46 | Carboxypeptidase |
| 44 | 4mdh_1 | 0.47 | 3chy_1 | 0.55 | Rossmann fold |
| 45 | 1min_1 | 0.47 | 2gbp_1 | 0.47 | |
| 46 | 1hmy_1 | 0.47 | 3chy_1 | 0.54 | |
| 47 | 3adk_1 | 0.49 | 3chy_1 | 0.53 | Adenylate kinase |
| 48 | 1etu_1 | 0.49 | 2gbp_1 | 0.49 | RAS/GTP-binding |
| 49 | 3enl_2 | 0.49 | 21bp_1 | 0.56 | TIM barrel |
| 50 | 3grs_2 | 0.49 | 3chy_1 | 0.66 | Glutathione reductase |
| 51 | 1ald_1 | 0.50 | 3chy_1 | 0.57 | TIM barrel |
| 52 | 3sc2_1 | 0.50 | 2gbp_2 | 0.66 | Hydrolase |
| 53 | 1gky_1 | 0.50 | 2gbp_1 | 0.50 | Adenylate kinase |
| 54 | 3pgm_1 | 0.50 | 21bp_2 | 0.67 | |
| 55 | 1pow_1 | 0.51 | 3chy_1 | 0.56 | |
| 56 | 1pow_2 | 0.51 | 1dri_1 | 0.53 | |
| 57 | 2taa_1 | 0.51 | 21bp_2 | 0.61 | TIM barrel |
| 58 | 1nip_1 | 0.51 | 1dri_1 | 0.63 | |
| 59 | 1pda_2 | 0.51 | 3chy_1 | 0.70 | |
| 60 | 1gp1_1 | 0.52 | 2gbp_1 | 0.52 | Glutathione peroxidase |
| 61 | 6acn_4 | 0.52 | 2gbp_1 | 0.52 | |
| 62 | 1btc_2 | 0.52 | 3chy_1 | 0.60 | |
| 63 | 1wsyb_1 | 0.52 | 3chy_1 | 0.59 | |
| 64 | 3pgk_1 | 0.52 | 1dri_1 | 0.79 | Rossmann fold |
| 65 | 1tim_1 | 0.52 | 2gbp_1 | 0.52 | TIM barrel |
| 66 | 2trx_1 | 0.53 | 1dri_1 | 0.63 | Glutathione peroxidase |
| 67 | 1pii_2 | 0.53 | 3chy_1 | 0.55 | TIM barrel |
| 68 | 2hhm_1 | 0.53 | 2gbp_1 | 0.53 | |
| 69 | 3pgk_2 | 0.53 | 2gbp_1 | 0.53 | Rossmann fold |
| 70 | 1alk_1 | 0.54 | 2gbp_1 | 0.54 | |
| 71 | 1ads_1 | 0.54 | 3chy_1 | 0.74 | TIM barrel |
| 72 | 1pii_1 | 0.54 | 2gbp_1 | 0.54 | TIM barrel |
| 73 | 1pda_1 | 0.55 | 1dri_1 | 0.56 | |
| 74 | 1pfk_2 | 0.55 | 21bp_2 | 0.72 | |
| 75 | 1rhd_2 | 0.55 | 1dri_2 | 0.73 | |
| 76 | 1trb_2 | 0.56 | 3chy_1 | 0.60 | Glutathione reductase |
| 77 | 3dfr_1 | 0.56 | 21bp_2 | 0.74 | Dihydrofolate reductase |
| 78 | 3fbp_2 | 0.57 | 1dri_1 | 0.58 | |
| 79 | 1btc_1 | 0.57 | 3chy_1 | 0.65 | TIM barrel |
| 80 | 1dsb_1 | 0.58 | 2gbp_1 | 0.58 | Glutathione peroxidase |
| 81 | 2yhx_2 | 0.60 | 21bp_1 | 0.72 | Actin |

**Table 3.** The 2gbp_1 SST family. Data is presented from both the single-center and multiple-center clustering schemes. The closest relative refers to the domain in the multiple family center with the lowest DSR score to the target domain. For the single-center method, all DSR values are with respect to 2gbp_1. The final column describes the fold family of the domain.

# Table 4

| | Family center | Number domains | Total number domains | Total domains (50%Len.) | Fraction coverage | Total fraction coverage | Description |
|---|---|---|---|---|---|---|---|
| 1 | 2lbp_2 | 86 | 86 | 68 | 56.62 | 56.62 | Extended Rossman fold family |
| 2 | 1tlk_1 | 35 | 121 | 93 | 23.70 | 80.32 | Immunoglobulin fold |
| 3 | 256b_1 | 20 | 141 | 104 | 10.80 | 91.52 | Four-helix bundle family |
| 4 | 1arb_2 | 9 | 150 | 113 | 7.57 | 99.09 | Mammalian serine protease family |
| 5 | 3cln_2 | 7 | 157 | 120 | 6.55 | 105.64 | Calcium-binding protein family |
| 6 | 5ldh_2 | 77 | 161 | 124 | 47.05 | 111.48 | Rossmann fold |
| 7 | 2tbvc_1 | 11 | 166 | 131 | 7.96 | 117.09 | Jellyroll fold |
| 8 | 1cpca_1 | 11 | 169 | 138 | 8.35 | 122.56 | Globin fold |
| 9 | 1mup_1 | 6 | 175 | 143 | 5.12 | 127.68 | Calycin (RBP) family |
| 10 | 1i1b_1 | 4 | 179 | 147 | 3.84 | 131.52 | Beta-trefoil fold |
| 11 | 2apr_2 | 4 | 183 | 151 | 3.55 | 135.07 | Aspartic acid protease family |
| 12 | 1trb_1 | 26 | 185 | 154 | 3.55 | 138.62 | Glutathione reductase family |
| 13 | 1gmp_1 | 5 | 189 | 157 | 3.61 | 142.16 | Barnase family |
| 14 | 1bgc_1 | 10 | 191 | 161 | 7.63 | 145.46 | Four-helix bundle, alternate connectivity |
| 15 | 1ald_1 | 17 | 192 | 167 | 10.27 | 148.61 | TIM barrel family |
| 16 | 1nsb_1 | 4 | 195 | 170 | 3.47 | 151.64 | Beta-propeller fold |
| 17 | 1bia_1 | 3 | 198 | 173 | 3.00 | 154.64 | Forked-head DNA-binding fold |
| 18 | 1tbp_1 | 5 | 203 | 175 | 2.74 | 157.38 | TATA-binding protein |
| 19 | 1cc5_1 | 3 | 206 | 178 | 2.63 | 160.01 | Cytochrome c family |
| 20 | 5cpa_1 | 16 | 206 | 182 | 9.68 | 162.43 | Carboxypeptidase family |
| 21 | 1pda_3 | 4 | 210 | 183 | 2.47 | 164.83 | Porphobilinogen deaminase, domain 3 |
| 22 | 1bov_1 | 3 | 213 | 185 | 2.35 | 167.18 | Staphylococcal nuclease family |
| 23 | 1mat_1 | 4 | 215 | 188 | 2.75 | 169.41 | Methionine aminopeptidase |
| 24 | 1aoz_1 | 17 | 215 | 188 | 12.02 | 171.46 | Immunoglobulin fold, copper-binding proteins |
| 25 | 2cpp_2 | 3 | 217 | 190 | 2.31 | 173.50 | Cytochrome P450, domain 2 |
| 26 | 1pii_1 | 19 | 217 | 192 | 11.16 | 175.51 | TIM barrel family |
| 27 | 1arp_1 | 2 | 219 | 194 | 2.00 | 177.51 | Cytochrome peroxidase, domain 1 |
| 28 | 1gcr_1 | 2 | 221 | 196 | 2.00 | 179.51 | Gamma-crystallin |
| 29 | 1lmb_1 | 3 | 223 | 198 | 2.32 | 181.51 | Lambda repressor family |
| 30 | 1onc_1 | 2 | 225 | 200 | 2.00 | 183.51 | Ribonuclease A family |
| 31 | 1ova_1 | 2 | 227 | 202 | 2.00 | 185.51 | Serpin family |
| 32 | 1plf_1 | 2 | 229 | 204 | 2.00 | 187.51 | Platelet factor 4 |
| 33 | 2pia_1 | 2 | 231 | 206 | 2.00 | 189.51 | Ubiquitin family |
| 34 | 4mdh_2 | 2 | 233 | 208 | 2.00 | 191.51 | Malate dehydrogenase, domain 2 |
| 35 | 1arb_1 | 8 | 234 | 209 | 5.80 | 193.36 | Mammalian serine protease family |
| 36 | 1arp_2 | 2 | 236 | 211 | 1.77 | 195.13 | Cytochrome peroxidase, domain 2 |
| 37 | 2cpp_1 | 2 | 238 | 213 | 1.76 | 196.90 | Cytochrome P450, domain 1 |
| 38 | 1atn_2 | 4 | 240 | 215 | 2.53 | 198.63 | Actin family |
| 39 | 1eaa_1 | 4 | 240 | 218 | 3.00 | 200.32 | Choramphenicol acetyltransferase |
| 40 | 1mns_1 | 3 | 241 | 220 | 2.35 | 201.91 | Mandelate racemase, domain 1 |

**Table 4.** The 40 SST families with the maximal coverage of the domain database. 'Number domains' refers to the number of domains possessing SSTs that belong to a given family. A running total of distinct domain hits is kept in the 'Total number domains' column. 'Total domains (50% Len.)' keeps track of the total number of different domains with SSTs spanning greater than 50% of the positions in the domain. 'Fraction coverage' refers to the sum of the proportion of each domain that is contained in a SST belonging to a given family. 'Total fraction coverage' is a running total of the fractional coverage, excluding double counting. The final column is a description of each family.

## Table 5

| | Domain | DSR | Alignment Length | DRMS | RMS |
|---|---|---|---|---|---|
| "Fingers" subdomain [1] | (1) 3cla_1 | 1.07 | 57 | 3.10 | 8.29 |
| "Palm" subdomain [2] | (1) 2gls_2 | 0.47 | 59 (67) | 1.40 (1.58) | 2.14 (2.22) |
| | (2) 2rus_1 | 0.58 | 59 (73) | 1.72 (2.19) | 2.43 (3.50) |
| | (3) 1bia_2 | 0.62 | 59 | 1.84 | 3.23 |
| "Thumb" subdomain [3] | (1) 1hmq_1 | 0.76 | 48 | 1.96 | 2.73 |
| | (2) 2tmv_1 | 0.88 | 48 | 2.26 | 3.72 |
| "Connection" subdomain [4] | (1) 1rnh_1 | 0.48 | 51 (66) | 1.29 (1.90) | 2.05 (3.12) |
| | (2) 1atn_2 | 0.51 | 51 (64) | 1.36 (1.74) | 2.07 (2.72) |
| | (3) 1lap_2 | 0.58 | 51 (62) | 1.56 (1.92) | 2.24 (2.98) |
| | (4) 1atn_1 | 0.59 | 51 | 1.57 | 2.54 |

**Table 5.** SSTs identified in the polymerase domain of HIV reverse transcriptase. For each of the four subdomains, the domains from the database with the lowest DSR scores are listed along with the minimum alignment length, DRMS, and RMS values. The comparison statistics for the standard alignment are presented in parenthesis.

# Chapter 3

# Protein Secondary Structure Prediction Using Nearest-Neighbor Methods

## 3.1 Abstract

We have studied the use of nearest-neighbor classifiers to predict the secondary structure of proteins. The nearest-neighbor rule states that a test instance is classified according to the classifications of "nearby" training examples from a database of known structures. In the context of secondary structure prediction, the test instances are windows of n consecutive residues, and the label is the secondary structure type ($\alpha$-helix, $\beta$-strand, or coil) of the center position of the window. To define the neighborhood of a test instance, we employed a novel similarity metric based on the local structural environment scoring scheme of Bowie et al. (1991). In this manner, we have attempted to exploit the underlying structural similarity between segments of different proteins to aid in the prediction of secondary structure. Furthermore, in addition to using neighborhoods of fixed radius, we explored a modification of the standard nearest-neighbor algorithm that involved defining an "effective radius" for each exemplar by measuring its performance on a training set. Using these ideas, we

achieved a peak prediction accuracy of 68%.

Finally, we sought to improve the biological utility of secondary structure prediction by identifying the subset of the predictions that are most likely to be correct. Toward this end, we developed a nearest-neighbor estimator that produced not the traditional "one-state" prediction ($\alpha$-helix, $\beta$-strand, or coil) but rather a probability distribution over the three states. It should be emphasized that this scheme estimates true probability values and that the resulting numbers are not pseudo-probability scores generated by simple normalization of the raw output of the predictor. Applying the mutual information statistic, we found that these probability triplets possess 58% more information than the one-state predictions. Furthermore, the probability estimates allow one to assign an a priori confidence level to the prediction at each residue. Using this approach, we found that the top 28% of the predictions were 86% accurate and the top 43% of the predictions were 81% accurate. These results indicate that, notwithstanding the limitations on overall accuracy of secondary structure prediction, a substantial proportion of a protein can be predicted with considerable accuracy.

## 3.2 Introduction

Predicting the secondary structure of proteins is an important intermediate step in the understanding of the tertiary structure of proteins. Secondary structure information can be incorporated into simulations that attempt to fold proteins. In addition, this information can be used to enhance the sensitivity of programs designed to identify proteins that are homologous to a query sequence.

In the field of artificial intelligence (AI), secondary structure prediction would be considered a typical classification problem: one attempts to predict the class (secondary structure) of a given instance based on its features (sequence). Many traditional AI classification methods have been used to predict secondary structure, including rule-based methods (Chou & Fasman, 1974), statistical methods (Gibrat et al., 1987; Stolarz et al., 1992), neural networks (Qian & Sejnowski, 1988; Holley & Karplus, 1989; Kneller et al., 1990), pattern-matching (Cohen et al., 1983; Rooman & Wodak, 1988), etc. Interestingly, all of the above achieved approximately the same level of proficiency: between 60 - 65% accuracy. It is believed that this plateau is the result of the inability of the current techniques to account for global interactions between segments of the protein separated by many residues.

Recently, there has been renewed interest in the AI community in methods that use a database of known examples to classify the test instance. These nearest-neighbor systems have achieved excellent results in a variety of problem domains (Aha et al., 1991). The basic idea of the nearest-neighbor approach is to use the labels of examples closely related to the test instance to determine the label of the test instance. Recently, several groups have successfully applied this methodology to the problem of protein secondary structure prediction (Nishikawa & Ooi, 1986; Levin et al., 1986; Zhang et al., 1992; Salzberg & Cost, 1992). Indeed, Zhang et al. (1992) found that their nearest-neighbor predictor (termed MBR) outperformed a neural network and a second-order Bayesian statistical method.

These recent studies on nearest-neighbor secondary structure prediction systems were somewhat limited by employing a narrow definition of "nearness" or "distance" based on sequence similarity. We sought to broaden this work in three ways – two of which are specific to nearest-neighbor systems while the third is applicable to almost any secondary structure predictor:

(1) We developed a hybrid scoring system that combined a sequence similarity matrix with the local structural environment scoring method of Bowie et al. (1991). Bowie and Eisenberg have demonstrated that by constructing a local environment profile based on the three-dimensional structure of a protein, one can identity distantly-related protein sequences likely to adopt the same structural fold as the starting protein. We reasoned that this same scoring system could be used to identify structurally similar segments in different proteins that possess low sequence similarity, thereby enhancing the sensitivity of a nearest-neighbor secondary structure predictor.

(2) We explored a generalization of the nearest-neighbor rule that involved establishing an "effective radius" for each exemplar. The radius of an exemplar determines whether or not it is involved in the classification of an instance. Based on its performance on a training set, an exemplar's radius can be shrunk or expanded.

(3) Finally, we explored the notion that, although it may not be possible to predict the entire secondary structure of a protein with high accuracy, it may be possible to identify a subset of the predictions (i.e., those residues) that are the most accurate. Toward this end, we developed a method to calculate a predicted probability distribution $P_i$ over the three possible states at residue $i$. This probability distribution represents a true probability estimate and is not a pseudo-probability score generated in an *ad hoc* fashion, e.g., by simply normalizing the raw outputs of a neural network. As we show, the probability triplets encode 58% more information than the simple one-state prediction and facilitate the identification of regions of high or low predictive confidence.

Using a neural network to combine the predictions from six different nearest-neighbor predictors that employed various combinations of two scoring sytems and three window sizes, we achieved a final prediction accuracy of 68%. This score is significantly higher than a perceptron neural network trained and tested over the same data set (63.4%). More importantly, by examining the estimated probability triplets, it was possible to distinguish the more reliable predictions. Interestingly, the 28% of the predictions assigned the highest level of confidence were 86% accurate, and the 43% of the predictions with the highest level of confidence were 81% accurate. These results indicate that a substantial portion of a protein sequence can be predicted with considerable accuracy.

## 3.3 Materials and Methods

### (a) Database

A database of 110 protein chains was selected from the Brookhaven Protein Data Bank. This list closely resembles the database of proteins assembled by Zhang et al. (1992). The secondary structure of these proteins was assigned using the program DSSP written by Kabsch and Sander (1983). Residues within $3_{10}$-helices were classified as coil. The sequences were less than 30% identical with each other with the exception of the pairs 1cse[i] and 2ci2[i] (35% identical) and 2abx[a] and 1nxb (43% identical). Percent sequence identity was determined by a standard dynamic programming algorithm (Needleman-Wunsch) using the Dayhoff PAM250 scoring matrix. The letter in the square brackets refers to the selected subunit. Below is a list of the protein chains used in this study (proteins in our database considered homologous to a given protein are listed in the curly braces):

(1) 155c, (2) 1abp {2gbp}, (3) 1acx, (4) 1cc5, (5) 1crn, (6) 1cse[i] {2ci2[i]}, (7) 1ctf, (8) 1cy3 {2cdv}, (9) 1ecd {1lh1, 1mbd, 2hhb[b]}, (10) 1etu, (11) 1fc1[a] {2fb4[h], 1mcp[l]}, (12) 1fc2[c], (13) 1fxb {4fd1}, (14) 1gcn, (15) 1gcr, (16) 1gd1[o], (17) 1gp1[a], (18) 1hip, (19) 1hmq[a], (20) 1i1b, (21) 1lh1 {1ecd, 1mbd, 2hhb[b]}, (22) 1lz1, (23) 1mbd {1ecd, 1lh1, 2hhb[b]}, (24) 1mcp[l] {2fb4[h], 1fc1[a]}, (25) 1mlt[a], (26) 1nxb {2abx[a]}, (27) 1paz {1pcy}, (28) 1pcy {1paz}, (29) 1pfk[a], (30) 1phh, (31) 1pp2[l], (32) 1ppt, (33) 1rhd, (34) 1rn3, (35) 1rnt, (36) 1sbt, (37) 1sgt {3rp2[a]}, (38) 1sn3, (39) 1tgs[i] {2ovo}, (40) 1tim[a], (41) 1ubq, (42) 256b[a], (43) 2aat, (44) 2abx[a] {1nxb}, (45) 2alp, (46) 2apr, (47) 2aza[b], (48) 2cab, (49) 2cro, (50) 2ccy[a], (51) 2cdv {1cy3}, (52) 2ci2[i] {1cse[i]}, (53) 2cpp, (54) 2cts, (55) 2cyp, (56) 2fb4[h] {1fc1[a], 1mcp[l]}, (57) 2gbp {1abp}, (58) 2gls[a], (59) 2gn5, (60) 2hhb[b] {1ecd, 1lh1, 1mbd}, (61) 2ins[a], (62) 2ins[d], (63) 2lbp, (64) 2lzm, (65) 2ovo {1tgs[i]}, (66) 2pab[a], (67) 2plv[a], (68) 2sns, (69) 2sod[b], (70) 2ssi, (71) 2stv, (72) 2taa, (73) 2tbv[c] {4sbv[c]}, (74) 351c, (75) 3adk, (76) 3b5c, (77) 3cla, (78) 3cln {3cpv, 3icb}, (79) 3cna, (80) 3cpv {3cln, 3icb}, (81) 3fxc, (82) 3fxn, (83) 3gap[a], (84) 3grs, (85) 3icb {3cln, 3cpv}, (86) 3icd, (87) 3pgk, (88) 3pgm, (89) 3rp2[a] {1sgt}, (90) 3rxn, (91) 3tln, (92) 3wrp, (93)

3xia, (94) 4cpa[i], (95) 4dfr[b], (96) 4fd1 {1fxb}, (97) 4mdh[a] {5ldh}, (98) 4pti, (99) 4sbv[c] {2tbv[c]}, (100) 4tsi[a], (101) 5cpa, (102) 5ldh {4mdh[a]}, (103) 6acn, (104) 7api[a], (105) 7wga[b], (106) 8adh, (107) 8atc[a], (108) 8atc[b], (109) 8cat[a], (110) 9pap.

There were 6148 helix positions (28.8%), 4123 beta-sheet positions (19.3%), and 11,078 coil positions (51.9%) in the database.

## (b) Training and testing procedure

We used a jackknife procedure to train and test the nearest-neighbor predictor (Kneller et al., 1990). Each protein in the database was successively chosen to be the test protein. The training set for a given test protein was constructed by removing the test protein and any homologous proteins from the database. In this manner, there was no overlap between the training data and the test data. Proteins considered homologous to a given protein based on close structural similarity are listed in braces (see above).

Two standard performance measures were used to assess prediction accuracy. $Q_3$ is the percentage of correct predictions,

$$Q_3 = \frac{q_\alpha + q_\beta + q_{coil}}{N} \cdot 100\% \tag{3.1}$$

where N is the total number of residues predicted, and $q_s$ is the number of residues of secondary structure type s that are predicted correctly. The Matthews' correlation coefficient is a more stringent measure of predictive accuracy,

$$C_s = \frac{(p_s \cdot n_s) - (u_s \cdot o_s)}{\sqrt{(n_s + u_s) \cdot (n_s + o_s) \cdot (p_s + u_s) \cdot (p_s + o_s)}} \tag{3.2}$$

where $p_s$ is the number of positive cases correctly predicted, $n_s$ is the number of negative cases correctly rejected, $o_s$ is the number of false positives, and $u_s$ is the number of false negatives (Schulz & Schirmer, 1979).

## (c) Nearest-neighbor method

The standard classification problem involves assigning a label from a set $L$ to any observation $\mathbf{X}$. The k nearest-neighbor rule classifies $\mathbf{X_i}$ based on the known labels of the $k$ nearest neighbors (exemplars) of $\mathbf{X_i}$. In this work, the instances were windows of $n$ consecutive amino acid residues (henceforth, n-segments) from the test protein, and the exemplars were n-segments of local structural environments and sequence derived from the training set proteins (see Fig. 1A). The n-segments were generated by collecting all overlapping windows of $n$ residues from the appropriate data set. The labels were the secondary structure type ($\alpha$-helix, $\beta$-sheet, or coil) of the center residue in the n-segment.

A given test instance was compared against all the exemplars using a scoring system, and the $k$ highest scoring exemplars (i.e., nearest neighbors) were noted. The raw output $\mathbf{Y_i}$ at position i is a triplet $(h_i, e_i, c_i)$ in which $h_i$, $e_i$, and $c_i$ respectively represent the number of nearest neighbors of structure type helix, sheet, and coil. The single state prediction at a residue is the label associated with a plurality of the $k$ high-scoring exemplars (see Figs. 1B and 1C). By examining the results from the training set, we established the rule that ties were broken in favor of helix followed by sheet.

## (d) Scoring system (distance metric)

Nearest-neighbor methods depend on a notion of "distance" between the test instances and the training exemplars. For this purpose, we adapted the local environment scoring method of Bowie and Eisenberg. Each protein in the training set was converted into a 3D structure profile by assigning each residue in the structure to an environment class. The local structural environment of a given residue was determined by three features: (i) the secondary structure; (ii) the solvent accessibility; (iii) the polarity. The three structural features were used in two different ways: either they were treated separately to construct three independent tables, or they were combined to create a single scoring table.

Secondary structure, as mentioned above, was assigned by the program DSSP. The fractional solvent accessibility of a given position was determined using the program DSSP to calculate the total solvent-exposed surface area of the residue in the protein and then dividing this number by the solvent-exposed surface area of the residue in a Gly-X-Gly tripeptide. The polarity of the environment was calculated by determining the fraction of the total surface area of a residue in contact with a polar atom, including the putative oxygen atoms of the surrounding solvent.

Environment classes were defined by systems of inequalities involving the solvent accessibility (denoted $a$) and polarity value (denoted $p$), with boundaries chosen to ensure approximately equally populated sets. Three different partitions of the $(a, p)$ values were used:

(1) $A_1 = \{0.0 \leq a < 0.12\}, A_2 = \{0.12 \leq a < 0.41\}, A_3 = \{0.41 \leq a \leq 1.0\}; P_1 = \{0.0 \leq p < 0.47\}, P_2 = \{0.47 \leq p < 0.63\}, P_3 = \{0.63 \leq p \leq 1.0\}$.

(2) $S_1 = \{0.0 \leq a < 0.12, 0.0 \leq p < 0.40\}, S_2 = \{0.0 \leq a < 0.12, 0.40 \leq p \leq 1.0\}, S_3 = \{0.12 \leq a < 0.41, 0.0 \leq p < 0.55\}, S_4 = \{0.12 \leq a < 0.41, 0.55 \leq p \leq 1.0\}, S_5 = \{0.41 \leq a \leq 1.0, 0.0 \leq p \leq 1.0\}$.

(3) $S_1 = \{0.0 \leq a < 0.12, 0.0 \leq p < 0.34\}, S_2 = \{0.0 \leq a < 0.12, 0.34 \leq p < 0.45\}, S_3 = \{0.0 \leq a < 0.12, 0.45 \leq p \leq 1.0\}, S_4 = \{0.12 \leq a < 0.41, 0.0 \leq p < 0.49\}, S_5 = \{0.12 \leq a < 0.41, 0.49 \leq p < 0.60\}, S_6 = \{0.12 \leq a < 0.41, 0.60 \leq p \leq 1.0\}, S_7 = \{0.41 \leq a \leq 1.0, 0.0 \leq p < 0.67\}, S_8 = \{0.41 \leq a \leq 1.0, 0.67 \leq p < 0.78\}, S_9 = \{0.41 \leq a \leq 1.0, 0.78 \leq p \leq 1.0\}$.

The score for matching a residue $R_i$ with a local structural environment $E_j$ was given by the information statistic:

$$SCORE(R_i, E_j) = \log_{10} \frac{P(R_i|E_j)}{P(R_i)} \tag{3.3}$$

where $P(R_i|E_j)$ is the probability of finding residue $i$ in environment $j$, and P$(R_i)$ is the probability of finding residue $i$ in any environment. The scores were determined by first removing the test protein from the data set and then collecting the necessary statistics from the remaining proteins. Thus, there were 110 scoring tables – one for each test protein.

### (e) Estimating the probability of each secondary structure type

It is important to know not only the one-state prediction of the most likely secondary structure type at a given position, but also an estimate of the probability associated with each type (i.e., the probability that position $i$ belongs to an $\alpha$-helix, $\beta$-sheet, or coil). To estimate the probability density for each prediction generated by our k nearest-neighbor predictor, we used, in turn, a k nearest-neighbor algorithm (Duda & Hart, 1973). Specifically, the output from any secondary structure predictor at position $i$ is a three-tuple vector $(Y_\alpha, Y_\beta, Y_c)$ – indicating the number of neighbors with label helix, beta-sheet, or coil. This triplet for the test instance was compared to the triplets for all instances in the training set (described in section (b)) using a Euclidean metric (i.e., $\sqrt{(\Delta h)^2 + (\Delta e)^2 + (\Delta c)^2}$ where $\Delta h$, $\Delta e$, and $\Delta c$ are the differences between the helix, sheet, and coil components of the raw output vectors), and the $k$ nearest neighbors were determined. The labels of the nearest neighbors were compiled and the proportion of the labels, $(\frac{h_i}{k}, \frac{e_i}{k}, \frac{c_i}{k})$, represented the estimate of the probability density at position $i$. For this purpose, we used a value of $k = 50$. Thus, this procedure measured how frequently a helix position, a beta-sheet position, and a coil position gave an output that is similar to $(Y_\alpha, Y_\beta, Y_c)$. Note that this method produced a true probability estimate, rather than merely a normalization of $(Y_\alpha, Y_\beta, Y_c)$.

### (f) Combining the ouputs from different predictors with a neural net

As described above in part (e), the outputs of the different nearest-neighbor predictors were converted into estimated probability scores using a $k$ nearest-neighbor algorithm. These probability values were the input to a neural net with no hidden layers (perceptron). The perceptron had a window size of 9. Supposing that there were 3 different predictions and given that each prediction is a 3-tuple vector and that the window size is 9, the perceptron would need to contain $\{[(3\cdot3)+1]\cdot9\} = 90$ input units. The one refers to the spacer unit. The database was divided into 4 groups; three groups were used to train the neural net and the fourth group was the test set.

## (g) Analysis of incorrect predictions

First, we determined percent accuracy for positions in each type of secondary structure. We then subdivided each secondary structure element into the ends and the middle positions and determined the performance on each. The first and last residues of each helix or $\beta$-strand were classified as end residues.

The correlation coefficient for correct (incorrect) predictions separated by d positions was calculated using the standard formula (Mendenhall et al., 1990):

$$r = \frac{\sum_i (y_i - \bar{y})(y_{i+d} - \bar{y})}{\sum_i (y_i - \bar{y})^2} \tag{3.4}$$

The data $y_i$ was either 1 or 0 depending on whether the prediction was correct, $d$ represents the separation distance, and $\bar{y}$ is the mean prediction accuracy (0.68).

## (h) Calculating the effective radius of each exemplar

In order to increase the influence of "reliable" exemplars relative to "unreliable" exemplars, we explored an approach that shifted the emphasis from the neighborhood of the test instances to the neighborhood of the training exemplars. The basic idea is to establish an effective radius for each exemplar. For a given test instance, an exemplar is counted as a neighbor of the test instance (i.e., is "active") if the test instance falls within the effective neighborhood of the exemplar. In the AI literature, this protocol has been termed the Alien Identification Rule (Dasarthy, 1991).

In practice, each exemplar was matched against a training set of instances. The radius r was specified in terms of the number of nearest training instances to the exemplar. For each exemplar $\mathbf{W_j}$, the radius determined a threshold score, $THRESH(\mathbf{W_j}, r)$. During the classification of a test instance $\mathbf{X_i}$, the set of "active" exemplars consisted of those exemplars for which the matching score was greater than the threshold score, $SCORE(\mathbf{X_i}, \mathbf{W_j}) \geq THRESH(\mathbf{W_j}, r)$. The training set for a given exemplar consisted of all instances in the database minus the test instance itself and any instances derived from proteins homologous to the protein from which the exemplar originated.

## (i) Information content of output

From an information theory perspective, the secondary structure of a protein can be viewed as a string composed of 3 letters: $h$, $e$, and $c$. We calculated the average amount of information per position in a typical secondary structure string using the formula for information entropy:

$$H(P) = -\sum_i p_i \cdot \log_2 p_i \tag{3.5}$$

where $p_i$ represents the probability that a given position in the string is of type $i$. For protein secondary structure, H(P) = 1.47 bits/position.

To measure the amount of information provided by both the standard one-state output and the probability triplets about the true secondary structure, we calculated the mutual information statistic (Stolarz et al., 1992):

$$I(S; D) = \sum_{i,j} Pr(S_i \cap D_j) \cdot \log_2 \frac{Pr(S_i \cap D_j)}{Pr(S_i) \cdot Pr(D_j)} \tag{3.6}$$

which is a standard measure of the amount of information provided by a prediction $D$ of a state $S$. For the standard one-state output, $Pr(D_j)$ is the probability (fraction) of predictions of type $D_j \in \{h, e, c\}$, $Pr(S_i)$ is the probability (fraction) of residues in the database that are in structure class $S_i \in \{h, e, c\}$, and $Pr(S_i \cap D_j)$ is the probability (fraction) of events in which a prediction of type $D_j$ is made at a position of type $S_i$. For the probability output encoding, each output triplet $\mathbf{P_k}$ was placed into a bin $B_j$ that encompassed a range of values that included $\mathbf{P_k}$. Then in the equation above, $Pr(D_j)$ is the probability that a prediction falls into bin $B_j$, and $Pr(S_i \cap D_j)$ is the probability that a prediction made at a position of type $S_i$ falls into bin $B_j$. There were 55 bins, and each bin spanned an interval of 10 percentage points for the helix and beta-strand probability values (e.g., $B_1 = \{\mathbf{P_k} | 0.0 \leq P_\alpha < 0.1, 0.0 \leq P_\beta < 0.1\}$).

In general, $I(S;D) \leq I(S;S) = H(S)$. That is, the mutual information of D about S is bounded above by the information of S about itself, with equality if and only if D is a perfect predictor.

## 3.4 Results

### (a) Scoring systems

The standard AI classification problem consists of assigning a class or label to a given test instance. The k nearest-neighbor rule classifies the test instance $\mathbf{X_i}$ based upon the known labels of the $k$ nearest neighbors of $\mathbf{X_i}$ chosen from a set of known exemplars. A key decision in any nearest-neighbor prediction scheme is the choice of the distance metric (scoring table) for relating the instances (the sample points to be classified) to the exemplars (the sample points whose labels are known). In the few previous studies employing nearest-neighbor classifiers for secondary structure prediction, the distance metric was based on sequence similarity (Levin et al., 1986; Zhang et al., 1992; Salzberg & Cost, 1992). In this paper, we sought to take advantage of the additional structural information inherent in the local structural environment scoring method of Bowie et al. (1991).

The local structural environment method involved assigning every residue of a protein with known three-dimensional structure to an "environment class" based on the local structural features of that position. In this manner, a 3D structure profile (environment sequence) was created that was converted into a set of training exemplars – n-segments of environment classes – by collecting all overlapping windows of n residues (see Fig. 1). Similarly, n-segments of sequence generated from the test protein became the test instances. Finally, a scoring table was set up that assigns a score for the alignment of each local environment class with each amino acid type. The score was simply the information value for pairing a residue $R_i$ with an environment $E_j$ (see Materials and Methods).

As originally described by Bowie and Eisenberg, the local structural environment of each position was determined by three local features: (i) secondary structure; (ii) solvent accessibility; (iii) fraction of contacts with polar atoms (polarity). In order to discretize the latter two features, the range of possible values was divided into three equally-sized categories. We wished to assess the relative importance of each structural feature in the classification process. Thus, instead of combining the different features into one table, we initially chose to construct three separate tables and then successively to add their contributions (Table 1, rows 1 - 3). Using a single table based on secondary structure alone resulted in a prediction accuracy of 59.2%. Adding the solvent accessibility table improved performance to 64.2%, and using all three tables produced a prediction accuracy of 65.1%. These results confirmed the importance of all three structural features. It should be noted that altering the exact ratio of the contributions of the three tables did not significantly affect the prediction (data not shown). Consequently, a 1:1:1 ratio was used.

We next wanted to know whether creating more specialized environment classes would be beneficial. We, therefore, combined the 3 local features into one table. Both a 15 state table (3 secondary structure types × 5 accessibility/polarity classes) and a 27 state table (3 secondary structure types × 9 accessibility/polarity classes) were tried. The performances of the single tables (65.1% and 65.6%, respectively) did not differ significantly from the performance of the combination of the three tables (see Table 1, rows 3, 6 and 7). In all cases, we were careful to construct the scoring table from a training set which consisted of all proteins in the database except for the test protein.

Finally, we wished to create a hybrid scoring system that combined the local environment scoring system with a scoring matrix based on sequence similarity. We selected for study the Dayhoff PAM250 matrix and a mutation matrix (Benner matrix) recently developed by Gonnet et al. (1992). Although neither matrix predicted well alone, 60.4% and 61.7% respectively (Table 1, rows 4 and 5), both slightly enhanced the performance of the local environment scoring tables by about 1% (Table 1, rows 8 and 9). In this hybrid system, the contributions of the two different scoring methods were approximately 1:1. The best scoring system, "Second/Access/Polar(15) + Benner", produced a prediction accuracy of 66.8% which is slightly higher than the predictive capability of a neural net (63.4%) without any hidden layers (perceptron) or the MBR predictor (64.1%), the nearest-neighbor prediction scheme developed by Zhang et al. (1992).

## (b) Number of nearest neighbors and window size

Another important parameter in any nearest-neighbor predictor is the number of nearest neighbors $k$ (i.e., size of the neighborhood) used to classify the test instance. Previous studies on nearest-neighbor classifiers of secondary structure employed values of $k$ between 1 and 25 (Zhang et al., 1992; Salzberg & Cost, 1992). Using a fixed window size of 13 residues and the "Second/Access/Polar(15) + Benner" scoring system, we tested a range of values for $k$ (see Table 2). As can be observed, the best classification results were achieved with $k$ between 50 and 200. Thus, as much as 1% of the database of approximately 20,000 exemplars were involved in the classification of each test instance.

We also tested the effect of window size on prediction accuracy. The simple-minded approach of capturing global interactions by enlarging the window size has not worked well for other classification methods. For example, Stolarz et al. (1992) found that a perceptron performed better with a window of 13 residues (63.5%) than a window of 25 residues (61.7%). The nearest-neighbor system operated best with $n = 19$, but achieved good results with window sizes as large as $n = 25$ or $n = 41$ (see Table 3). It is possible that the decline in accuracy with very large windows can be attributed to the failure to allow gaps or insertions during the alignment of the exemplars with the test instance.

## (c) Combining several predictions with a neural network

We noticed that different scoring systems and different window sizes in the nearest-neighbor classifier produced significantly different predictions. We reasoned that combining information from these different predictions could improve performance. Indeed, Zhang et al. (1992) have demonstrated that using a neural network to combine the outputs of several different classification methods resulted in a classification accuracy superior to that of the individual predictions. Employing two scoring systems and three window sizes, we collected six different outputs. These outputs were converted into estimated probability values with a k nearest-neighbor algorithm described below (see section (f) or Materials and Methods). A standard testing and training procedure was used with a neural network with no hidden layers (perceptron) possessing a window size of 9. We first merged the outputs from the 3 window sizes for each scoring system and then the resulting two predictions were combined to produce the final prediction (see Table 4).

The overall prediction accuracy was 68.0%. The correlation coefficients for the three types of secondary structure were $C_\alpha = 0.52$, $C_\beta = 0.41$, and $C_{coil} = 0.44$. The variation in prediction accuracy between proteins was quite large ranging from 100.0% (1ppt) to 44.7% (2pab). A full list of the prediction accuracy for all 110 proteins is provided in Table 5. Part of this variability could be attributed to differences in the content of $\beta$-sheet in the proteins. Indeed, we found that all $\alpha$-helical proteins (70.9%) were better predicted than $\alpha/\beta$ proteins (67.9%) or all $\beta$-sheet (65.6%) proteins.

Table 6 compares the reported accuracy for a variety of prediction schemes. Because different authors employ somewhat different databases and training/testing protocols, direct comparison with previously published work is problematic. We did, however, directly test the memory-based reasoning (MBR) method of Zhang et al. (1992) and a perceptron neural network on our database (see Table 1): our nearest-neighbor method performed significantly better in this direct comparison (at the 0.99 confidence level, see equation (6) of Zhang et al., 1992), yielding 68% accuracy as against 64.1% and 63.4% for the other methods. We also note that neural networks trained and tested by other investigators resulted in prediction accuracies similar to our perceptron data (Qian & Sejnowski, 64.3%; Holley & Karplus, 63.2%; Kneller et al., 65%, Zhang et al., 63.1%), suggesting that differences in the databases did not have a dramatic effect on performance.

## (d) An analysis of incorrect predictions

We next analyzed the location and distribution of the prediction errors. As shown in Table 7, we decomposed prediction accuracy into secondary structure type and relative location in the secondary structure element. Table 7, part A, demonstrates that $\beta$-strand positions were predicted much more poorly than the helix or coil positions: only 41.7% of all $\beta$-strand residues were predicted correctly. Furthermore, we found that the ends of the sheets and helices were more susceptible to errors than the middle positions. Indeed, as shown in Table 7, part B, the prediction accuracy at the ends of helices was 36.8%, whereas the prediction accuracy at the central residues of helices was 68.6%. (N.B. Some of these errors may be due to ambiguity in defining the precise boundaries of secondary structure elements).

Finally, we wished to address the question of whether the errors were uniformly distributed or occurred in clusters. One simple indicator is provided by measuring the length of an average run of incorrect predictions. If errors were uniformly distributed (with 68% overall accuracy), one would expect that correct predictions would occur with an average run length of 3.09 residues and incorrect predictions would occur with an average run length of 1.47 residues. In fact, our nearest-neighbor method produced runs of correct and incorrect predictions of 6.77 and 3.35 residues, respectively. This indicates a non-random clustering of errors. This point is illustrated in more detail in Figure 2, which shows the correlation between the correctness (incorrectness) of the predictions at positions $i$ and $i + d$, for various separation distances $d$. This autocorrelation is over 0.50 at a distance of 1 and is still substantial even at a distance of 5. This clustering of errors may explain why our method and those of others (Zhang et al., 1992) sometimes miss entire secondary structure elements.

## (e) Calculating the radius of each exemplar

Because the exemplars appeared to vary in reliability, we decided to study a variation of the k nearest-neighbor rule, termed the Alien Identification Rule (Dasarthy, 1991), that focuses on the neighborhood around each exemplar rather than the neighborhood around each instance. The standard protocol for nearest-neighbor prediction is an *instance-based* approach: an exemplar $\mathbf{W_j}$ participates in the classification of an instance $\mathbf{X_i}$ if it is one of the $k$ nearest neighbors to $\mathbf{X_i}$. Alternatively, one can use an *exemplar-based* approach: one defines a "sphere of influence" with a radius $r$ for each exemplar $\mathbf{W_j}$, and declares that a given exemplar $\mathbf{W_j}$ is "active" in the classification of instance $\mathbf{X_i}$ if $\mathbf{X_i}$ is located within the radius $r$ of $\mathbf{W_j}$. The potential advantage of this approach is that "unreliable" exemplars (defined by some criteria) can be given a small radius, while reliable examplars can be assigned a larger radius. The approach is similar in certain respects to that adopted by Levin et al. (1986) in which a training exemplar participated in the classification of a test instance if the two sequence segments contained more than a prespecified number of identities.

Using this approach, each test instance was compared against the database of exemplars as was usually done. Then, any exemplars with a matching score that was below a certain cutoff, specific to each exemplar, were removed. The remaining exemplars were used to classify the test instance. For each exemplar, the cutoff score was determined by comparing the exemplar against a set of training instances. The highest-scoring $r$ training instances specified the cutoff score, $THRESH(\mathbf{W_j}, r)$, above which the top $r$ instances scored. Thus, the test instance $\mathbf{X_i}$ fell within the "sphere of influence" of $\mathbf{W_j}$ if the matching score, $SCORE(\mathbf{X_i}, \mathbf{W_j})$, was greater than or equal to $THRESH(\mathbf{W_j}, r)$. With a value of $r = 50$, this approach achieved a level of performance (66.7%; Table 8, row 1) comparable to that of the standard instance-based method (66.8%; Table 8, row 8).

One advantage of the exemplar-based system is that it is possible to compile statistics measuring the performance of each exemplar at a given radius – allowing one to identify good and bad exemplars. Figure 3 presents a picture of the overall predictive performance of the exemplars when $r = 50$. Although the majority of the exemplars (61%) had a predictive accuracy between 50% and 80%, there were a considerable number that fell far above or below this region. Further attention was devoted to these outlying exemplars.

We attempted to improve the predictions by decreasing the radius of the most unreliable exemplars to 0 and expanding the radius of the most reliable exemplars. This strategy has improved the performance of nearest-neighbor classifiers in several cases (Aha et al., 1991). We began by removing the poorest-performing 10%, 20%, or 30% of the exemplars. In order to avoid the inclusion of test data into the performance statistics, the statistics for each exemplar were generated from a training set (see Materials and Methods). As can be seen in Table 8, elimination of the unreliable exemplars resulted in almost no improvement in prediction. We then attempted to couple the pruning process with the expansion of good exemplars by removing the worst $X\%$ of the exemplars and then expanding the radius of the top $X\%$ of the exemplars from 50 to 100 where $X = 10$, 20, and 30 (rows 5, 6, and 7, Table 8). A slight but noticeable gain (0.6%) in performance was observed.

### (f) Estimating the probabilities of each secondary structure type

Traditionally, the final output of a secondary structure predictor is a "one-state" prediction of the most likely secondary structure type at each position. We reasoned that it might be more informative to predict a probability distribution $\mathbf{P_i}$ over the three possible states at each residue $i$.

A variety of methods exist for estimating probability distributions including Parzen Window techniques, k nearest-neighbor algorithms, and kernel estimators (Duda & Hart, 1973). We selected for study the k nearest-neighbor algorithm because of its simplicity and robustness. The basic idea is as follows. The raw output of our nearest-neighbor secondary structure predictor was a list of triplets $\mathbf{Y_i} = (h_i, e_i, c_i)$, where $h_i$, $e_i$, and $c_i$ are the number of nearby exemplars that were in the helix, strand, or coil state, respectively, for position i. In order to obtain instead a probability distribution $\mathbf{P_i} = \mathbf{P}(\mathbf{Y_i})$, we identified the training examples which yielded an output vector near to $\mathbf{Y_i}$ and defined $\mathbf{P_i}$, based on the empirical probability distribution observed for these nearby training examples (see Materials and Methods). Thus, we measured the proportion of helix positions, beta-sheet positions, and coil positions that had an output similar to $\mathbf{Y_i}$. In this application, we defined distance between output vectors $\mathbf{Y_i}$ by the standard Euclidean distance metric, and we used a value of $k = 50$. It should be noted that this probability density estimation method estimates true probability values and not ad hoc pseudo-probability scores generated by normalization of the raw output $\mathbf{Y_i}$.

It was important to assess the accuracy of the predicted probability distributions.

To this end, we grouped probability values into bins spanning an interval of 10 percentage points and measured the level of accuracy of the predictions associated with each interval. Parts A and B of Table 9 show the relationship between the *predicted* accuracy and the *actual* accuracy. The results indicate that the predicted probability does not deviate substantially from the actual.

The output of estimated probability triplets provides a way to discriminate between positions of high predictive confidence and positions of low confidence. Indeed, looking at Table 9 (part A), one observes that the predictions with predicted accuracy greater than 80% represented 28.4% of all predictions and were 85.9% accurate, while those with predicted accuracy greater than 70% represented 43.3% of all predictions and were 81.3% accurate. As might be expected from the large number of coil positions in the database, the coil predictions tended to have higher probability values than the beta-sheet or helix predictions. Indeed, Table 10 shows that 34.3% of all coil predictions had a probability value greater than 0.8 while only 24.4% of helix predictions and 11.6% of beta-sheet predictions achieved a similar level of confidence.

The primary motivation for converting the output of the predictor into probability triplets was to increase the information content of the predictions. We compared the information content of the two different output descriptions (one-state vs. probability distribution) using the mutual information statistic. From an information theory perspective, the secondary structure of a protein can be viewed as a string composed from a three-letter alphabet ($h$, $e$, and $c$). We used the formula for information entropy to calculate the average amount of information per position in a typical protein secondary structure:

$$H(P) = -\sum_{i} p_i \cdot \log_2 p_i \tag{3.5}$$

where $p_i$ represents the probability of each type of secondary structure. We found that protein secondary structure contains 1.47 bits/position of information (a binary number contains one bit of information per position). We then determined the information provided by the predictions relative to the actual secondary structure using the following formula for mutual information:

$$I(S;D) = \sum_{i,j} Pr(S_i \cap D_j) \cdot \log_2 \frac{Pr(S_i \cap D_j)}{Pr(S_i) \cdot Pr(D_j)} \tag{3.6}$$

where $Pr(D_j)$ is the probability of a prediction of type $D_j$, $Pr(S_i)$ is the probability of a residue in structure class $S_i$, and $Pr(S_i \cap D_j)$ is the probability of events in which a prediction of type $D_j$ is made at a position of type $S_i$. We found that the output consisting of the estimated probability triplets possessed 0.41 bits/position of information whereas the traditional one-state output contained 0.28 bits/position. Thus, the estimated probability output contained 58% more information than the one-state output. Nonetheless, even the probability triplets captured less than $(\frac{0.41}{1.47})$ or 28% of the total amount of information present in the secondary structure.

## 3.5 Discussion

We investigated the application of a nearest-neighbor classification system to the problem of secondary structure prediction. Our work introduced two innovations to this approach. First, we implemented a hybrid scoring system that combined the Bowie and Eisenberg local environment scoring system with a sequence similarity matrix (Benner matrix). This scoring table was designed to detect subtle structural similarities between segments of different proteins. Indeed, it is encouraging that this method performs better than any previous single prediction method. In addition, when six different predictions were combined with a neural network, a prediction accuracy of 68% was observed. We experimented with other scoring tables including those based on other structural features such as hydrophobic moment, size, and local conformation. None resulted in any improvement. We also attempted to merge our scoring system with the value-difference metric employed by the MBR predictor. Again, there was no increase in predictive accuracy (data not shown).

Secondly, we studied the concept of establishing the effective radius of each exemplar. The radius of the exemplar was set by determining a threshold score, $THRESH(\mathbf{W_j}, r)$, for each exemplar. Using a radius of 50, this exemplar-based approach performed comparably to the standard instance-based scheme. Pruning unreliable exemplars by setting their radius to be 0 and expanding the radius of reliable exemplars from 50 to 100 using performance statistics generated from a training set resulted in a 0.6% improvement in prediction. Although this methodology produced only a modest improvement in prediction accuracy, it may be the approach of choice in the future because of its greater flexibility. Different types of exemplars could be accommodated within the same predictor. For example, two exemplars could possess different window sizes and use different scoring systems and yet be used to classify the same test instance. Unlike the standard instance-based method, the scores between a given test instance and different exemplars are not directly compared.

The merits of the nearest-neighbor approach to secondary structure prediction include simplicity, flexibility, and a straightforward interpretation. When applied to the protein secondary structure problem, the nearest-neighbor rule captures the biologically appealing idea of using structural homology between protein segments to aid in prediction. It is possible that the slight improvement in overall performance observed with our method can be attributed to the ability of the Bowie and Eisenberg scoring system to detect subtle structural relationships between parts of different proteins which possess little sequence similarity. One possible area of research is enhancing the sensitivity of methods designed to detect this underlying structural homology.

Ultimately, the key to a dramatic increase in prediction accuracy is the incorporation of global, long-range information into the prediction scheme. Using a larger window size is one attempt in this direction. Unfortunately, nearest-neighbor methods suffer from the same defect as other approaches and exhibit a decline in performance with increasing window size. Part of the problem may be the failure to allow gaps or insertions during the alignment of the test instance to the training exemplars. We are currently investigating using a dynamic programming algorithm that permits gaps and insertions during the matching phase of the procedure.

Finally, we wished to study the advantages of transforming the raw output of our predictor into estimated probabilities of each type of secondary structure. We used a k nearest-neighbor algorithm to estimate a predicted probability distribution at each residue. It should be emphasized that this algorithm is not limited to nearest-neighbor predictors, but can be applied to the output of any secondary structure predictor. These predicted probability distributions had the advantage that they contained 58% more information than the standard one-state predictions, 0.41 versus 0.26 bits/position. However, since the typical protein secondary structure contains on average 1.47 bits/position of information, this still represents less than 28% of the information present in protein secondary structure.

Moreover, the estimated probability triplets allow one to identify *a priori* the residues that are most likely to be accurately predicted. We found that the top 28.4% of the predictions were 85.9% accurate, and the top 43.3% of the predictions were 81.3% accurate. Others have previously formulated the concept of confidence scales (Biou et al., 1988) or prediction strength (Holley & Karplus, 1989), but their confidence values were calculated *a posteriori* after the predictions had been specified and evaluated. In other words, because the confidence values were not derived using appropriate training and testing procedures one cannot be certain about the accuracy of these values on a new protein. Furthermore, no attempt was made to determine the probability values for all three secondary structure types at each position.

The approach of calculating a predicted accuracy at each residue offers a way to increase the biological utility of secondary structure prediction, notwithstanding the fact that the overall predictive accuracy seems to have reached a plateau at slightly less than 70%. We suggest that secondary structure predictions and methods should henceforth include predictions of the probability distribution (rather than simply one-state predictions) in order to assist consumers of such predictions in discriminating confident assertions from wild guesses. In Figure 4, we illustrate a sample output for the first 25 residues of cytochrome c550 (155c) from our prediction scheme which lists both the final one-state output and the associated probabilities for each secondary structure type. Although our work produced only a slight improvement in overall prediction accuracy compared to other prediction methods, there was a significant increase in the total information provided by the predictor.

# References

[Aha et al., 1991] Aha, D. W., Kibler, D., and Albert, M. A. (1991). Instance-based learning algorithms. *Machine Learning*, 6:37–66.

[Biou et al., 1988] Biou, V., Gibrat, J. F., Levin, J. M., Robson, B., and Garnier, J. (1988). Secondary structure prediction: Combination of three different methods. *Protein Eng.*, 2:185–191.

[Bowie et al., 1991] Bowie, J. U., Luthy, R., and Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, 253:164–170.

[Chou and Fasman, 1974] Chou, P. Y. and Fasman, G. D. (1974). Prediction of protein conformation. *Biochemistry*, 13:222–244.

[Cohen et al., 1983] Cohen, F. E., Abarbanel, R. A., Kuntz, I. D., and Fletterick, R. J. (1983). A combinatorial approach to secondary structure prediction: $\alpha/\beta$ proteins. *Biochemistry*, 22:4894–4904.

[Dasarthy, 1991] Dasarthy, B. V. (1991). *Nearest Neighbor Norms: NN Pattern Classification Techniques.* IEEE Computer Society Press, Los Alamitos.

[Duda and Hart, 1973] Duda, R. O. and Hart, P. E. (1973). *Pattern Classification and Scene Analysis.* John Wiley & Sons, New York.

[Gibrat et al., 1987] Gibrat, J. F., Garnier, J., and Robson, B. (1987). Further developments of protein secondary structure prediction using information theory. *J. Mol. Biol.*, 198:425–443.

[Gonnet et al., 1992] Gonnet, G. H., Cohen, M. A., and Benner, S. A. (1992). Exhaustive matching of the entire protein sequence database. *Science*, 256:1443–1445.

[Holley and Karplus, 1989] Holley, L. and Karplus, M. (1989). Protein secondary structure prediction with a neural network. *Proc. Natl. Acad. Sci., U.S.A.*, 86:152–156.

[Kabsch and Sander, 1983] Kabsch, W. and Sander, C. (1983). Dictionary of protein secondary structures: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22:2577–2637.

[Kneller et al., 1990] Kneller, D. G., Cohen, F. E., and Langridge, R. (1990). Improvements in protein secondary structure prediction by an enhanced neural network. *J. Mol. Biol.*, 214:171–182.

[Levin et al., 1986] Levin, J. M., Robson, B., and Garnier, J. (1986). An algorithm for secondary structure determination in proteins based on sequence similarity. *FEBS Letters*, 205:303–308.

[Mendenhall et al., 1990] Mendenhall, W., Wackerly, D. D., and Scheaffer, R. L. (1990). *Mathematical Statistics with Applications.* PWS-Kent, Boston.

[Nishikawa and Ooi, 1986] Nishikawa, K. and Ooi, T. (1986). Amino acid sequence homology applied to the prediction of protein secondary structures, and joint prediction with existing methods. *Biochimica et Biophysica Acta*, 871:45–54.

[Qian and Sejnowski, 1988] Qian, N. and Sejnowski, T. J. (1988). Predicting the secondary structure of globular proteins using neural network models. *J. Mol. Biol.*, 202:865–884.

[Rooman and Wodak, 1988] Rooman, M. J. and Wodak, S. J. (1988). Identification of predictive sequence motifs limited by protein structure database size. *Nature*, 335:45–49.

[Salzberg and Cost, 1992] Salzberg, S. and Cost, S. (1992). Predicting protein secondary structure with a nearest-neighbor algorithm. *J. Mol. Biol.*, 227:371–374.

[Schulz and Schirmer, 1979] Schulz, G. E. and Schirmer, R. H. (1979). *Principles of Protein Structure*. Springer-Verlag, New York.

[Stolarz et al., 1992] Stolarz, P., Lapedes, A., and Xia, Y. (1992). Predicting protein secondary structure using neural net and statistical methods. *J. Mol. Biol.*, 225:363–377.

[Zhang et al., 1992] Zhang, X., Mesirov, J. P., and Waltz, D. L. (1992). Hybrid system for protein secondary structure prediction. *J. Mol. Biol.*, 225:1049–1063.

## 3.6 Figures and Tables

## A. Creation of Training and Test Sets (window = 13).

| Test Protein | | Training Protein | |
|---|---|---|---|
| NEGDAAKGEKEFNKCK... | (sequence) | FF35523542552254... | (3D structure profile) |
| 1. ------NEGDAAK | Test | 1. ------FF35523 | Training |
| 2. -----NEGDAAKG | Instances | 2. -----FF355235 | Exemplars |
| 3. ----NEGDAAKGE | | 3. ----FF3552354 | |
| 4. ---NEGDAAKGEK | | 4. ---FF35523542 | |
| . | | . | |
| . | | . | |
| . | | . | |

## B. Identification of Nearest Neighbors.

### 1) Matching:

| Instance | Exemplars | Matching Score |
|---|---|---|
| 1. ------NEGDAAK | 1. ------FF35523 | 156 |
| | 2. -----FF355235 | 156 |
| | 3. ----FF3552354 | 7 |
| | . | |
| | . | |
| | . | |

### 2) Ordering (k=50): (for test instance #1)

| Exemplar No. | | Matching Score | SS type of exemplar |
|---|---|---|---|
| 1. | 7383 | 290 | C |
| 2. | 18033 | 260 | C |
| . | . | . | . |
| . | . | . | . |
| . | . | . | . |
| 50. | 19796 | 176 | C |

## C. Final Prediction.

(for test instance #1)

1) $Y_1$ = (0,0,50); H = 0, E = 0, C = 50.

Figure 1

**Figure 1**. A schematic description of the nearest-neighbor method applied to secondary structure prediction. Outlined is the procedure used to create the set of test instances and training exemplars (part A), to identify the nearest neighbors of a given test instance (part B), and to tabulate the final prediction (part C). The 15 environment classes for the 3D structure profile are represented using hexadecimal notation. A more detailed description of this protocol and an explanation of how the scoring table was constructed are provided in the Materials and Methods section.

## Figure 2

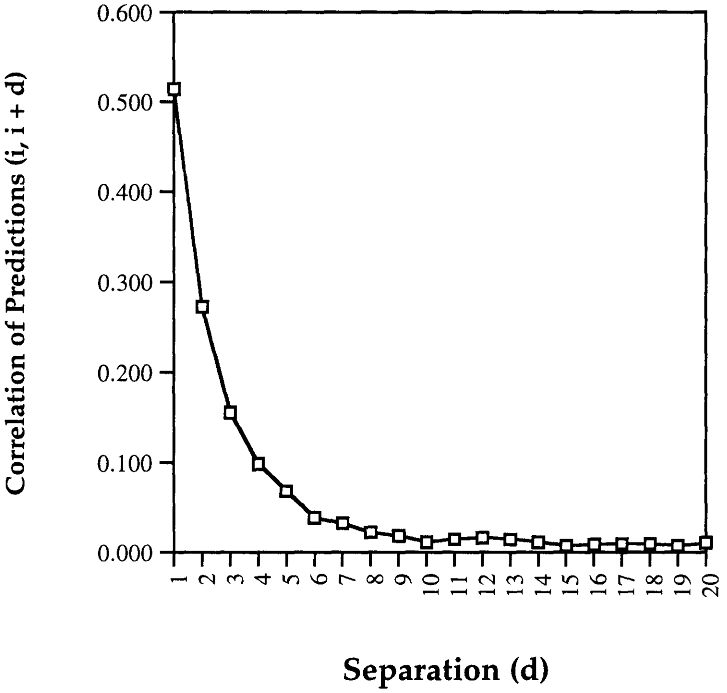


**Figure 2.** Correlation coefficients of correct (or incorrect) predictions separated by d positions. The plot is identical whether correct or incorrect predictions are analyzed. The standard formula for statistical correlation was used (see Materials and Methods).

Figure 3

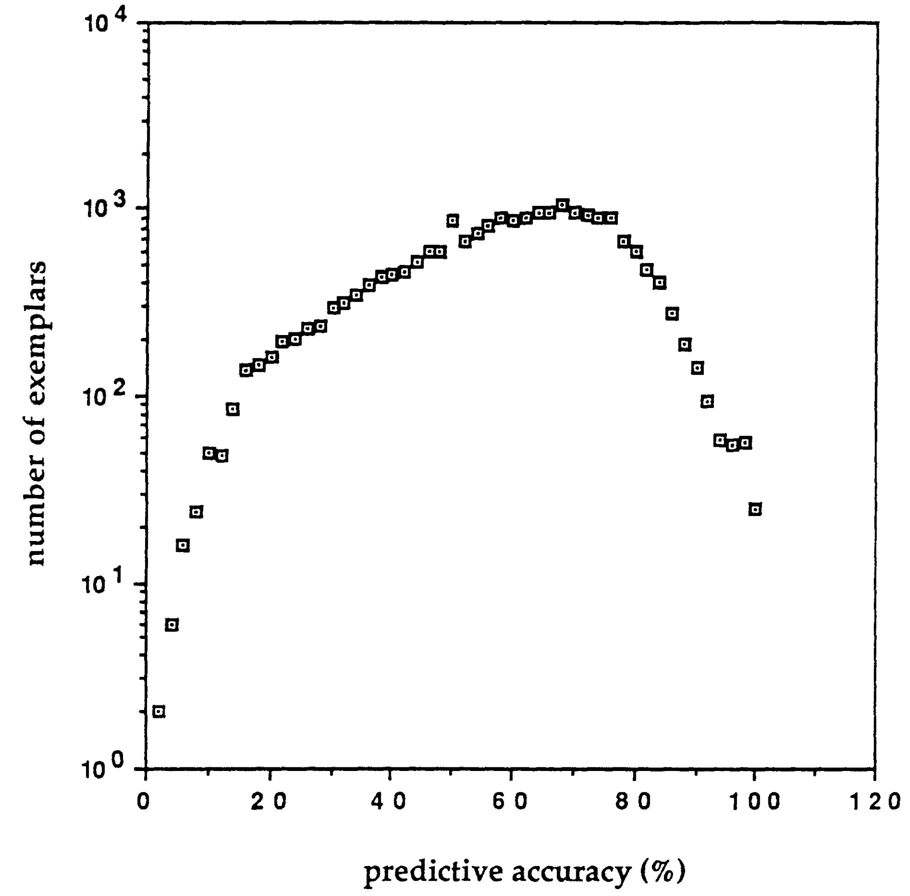


**Figure 3.** A graph depicting the performance characteristics of the exemplars against the whole database. The number of exemplars possessing each level of predictive accuracy is plotted. The scoring table was "Second/Access/Polar(15) + Benner", the radius r = 50, and the window size was 13.

PROTEIN = 155c

| Position | One-state prediction | Probability values |  |  | True secondary structure |
|---|---|---|---|---|---|
|  |  | $P_{\alpha}$ | $P_{\beta}$ | $P_{coil}$ |  |
| 1 | c | 0.00 | 0.10 | 0.90 | c |
| 2 | c | 0.04 | 0.00 | 0.96 | c |
| 3 | c | 0.04 | 0.02 | 0.94 | c |
| 4 | c | 0.02 | 0.04 | 0.94 | c |
| 5 | c | 0.10 | 0.08 | 0.82 | c |
| 6 | c | 0.34 | 0.02 | 0.64 | h |
| 7 | c | 0.24 | 0.10 | 0.66 | h |
| 8 | c | 0.46 | 0.02 | 0.52 | h |
| 9 | h | 0.46 | 0.10 | 0.44 | h |
| 10 | h | 0.50 | 0.06 | 0.44 | h |
| 11 | h | 0.48 | 0.08 | 0.44 | h |
| 12 | h | 0.48 | 0.04 | 0.48 | h |
| 13 | c | 0.40 | 0.06 | 0.54 | c |
| 14 | c | 0.38 | 0.00 | 0.62 | c |
| 15 | c | 0.24 | 0.04 | 0.72 | c |
| 16 | c | 0.18 | 0.08 | 0.74 | c |
| 17 | c | 0.20 | 0.22 | 0.58 | c |
| 18 | c | 0.18 | 0.20 | 0.62 | c |
| 19 | c | 0.14 | 0.40 | 0.46 | e |
| 20 | e | 0.24 | 0.60 | 0.16 | e |
| 21 | e | 0.14 | 0.46 | 0.40 | c |
| 22 | c | 0.16 | 0.40 | 0.44 | c |
| 23 | c | 0.08 | 0.16 | 0.76 | c |
| 24 | c | 0.02 | 0.08 | 0.90 | c |
| 25 | c | 0.00 | 0.06 | 0.94 | c |

**Figure 4.** A sample output for the first 25 positions of cytochrome c550 (155c). The standard one-letter abbreviations for the three types of secondary structure is used: 'h'= helix, 'e' = beta-strand, and 'c' = coil. $P_{\alpha}$, $P_{\beta}$, and $P_{coil}$ represent the estimated probabilities for each type of secondary structure at a given position.

## Table 1

The performances of different scoring tables

| Scoring Matrix | Accuracy (%) |
|---|---|
| 1. Second | 59.2 |
| 2. Second + Access | 64.2 |
| 3. Second + Access + Polar | 65.1 |
| 4. Dayhoff | 60.4 |
| 5. Benner | 61.7 |
| 6. Second/Access/Polar(15) | 65.1 |
| 7. Second/Access/Polar(27) | 65.6 |
| 8. Second + Access + Polar + Benner | 66.1 |
| 9. Second/Access/Polar (15) + Benner | 66.8 |
| 10. Neural Net (Perceptron) | 63.4 |
| 11. MBR | 64.1 |

Window size = 13; k (the number of nearest neighbors) = 50.

"Second + Access + Polar" indicates 3 separate tables.

"Second/Access/Polar(15)" refers to a single table with the number of states specified by the number in the parentheses.

## Table 2

The effect of the number of nearest neighbors on prediction

| Number of Nearest Neighbors | Accuracy (%) |
|---|---|
| 1 | 58.3 |
| 10 | 65.4 |
| 25 | 66.1 |
| 50 | 66.8 |
| 100 | 66.9 |
| 200 | 66.8 |
| 400 | 66.3 |

Window size = 13; scoring table = "Second/Access/Polar(15) + Benner".

## Table 3

The effect of window size on prediction

| Window Size | Accuracy (%) |
|---|---|
| 7 | 63.4 |
| 13 | 66.8 |
| 19 | 67.1 |
| 25 | 66.5 |
| 41 | 65.0 |

Scoring table = "Second/Access/Polar(15) + Benner"; k=50.

Table 4

Combining multiple predictions with a neural network

| Scoring Table | Window Size | Accuracy (%) | | |
|---|---|---|---|---|
| Second/Access/Polar(15)+ Benner | 13 | 66.8 | 67.8 | 68.0 |
| | 19 | 67.1 | | |
| | 25 | 66.5 | | |
| Second+Access+Polar+ Benner | 13 | 66.1 | 67.4 | |
| | 19 | 66.1 | | |
| | 25 | 65.6 | | |

The predictions from the 3 different window sizes for each scoring table were first combined with a neural net (perceptron). The resulting two outputs were then fed into a second neural net to produce the final prediction. A description of the procedure used to train and test the neural nets is provided in the Materials and Methods section.

# Table 5 Prediction accuracy for each protein

| PROTEIN | ACCURACY (%) |
|---|---|
| 1 155c | 71.9 |
| 2 1abp | 66.3 |
| 3 1acx | 64.5 |
| 4 1cc5 | 65.1 |
| 5 1crn | 50.0 |
| 6 1cse[i] | 69.8 |
| 7 1ctf | 55.9 |
| 8 1cy3 | 72.0 |
| 9 1ecd | 60.3 |
| 10 1etu | 69.5 |
| 11 1fc1[a] | 56.8 |
| 12 1fc2[c] | 86.0 |
| 13 1fxb | 79.0 |
| 14 1gcn | 69.0 |
| 15 1gcr | 60.9 |
| 16 1gd1o | 64.1 |
| 17 1gp1[a] | 76.1 |
| 18 1hip | 82.4 |
| 19 1hmq[a] | 53.1 |
| 20 1i1b | 72.2 |
| 21 1lh1 | 63.4 |
| 22 1lz1 | 68.5 |
| 23 1mbd | 70.6 |
| 24 1mcp[l] | 64.1 |
| 25 1mlt[a] | 73.1 |
| 26 1nxb | 66.1 |
| 27 1paz | 62.5 |
| 28 1pcy | 69.7 |
| 29 1pfk[a] | 74.7 |
| 30 1phh | 63.2 |
| 31 1pp2[l] | 59.8 |
| 32 1ppt | 100.0 |
| 33 1rhd | 69.3 |
| 34 1rn3 | 65.3 |
| 35 1rnt | 63.5 |
| 36 1sbt | 65.5 |
| 37 1sgt | 68.2 |
| 38 1sn3 | 63.1 |
| 39 1tgs[i] | 66.1 |
| 40 1tim[a] | 72.9 |
| 41 1ubq | 80.3 |
| 42 256b[a] | 89.6 |
| 43 2aat | 69.9 |
| 44 2abx[a] | 86.5 |
| 45 2alp | 51.5 |
| 46 2apr | 64.6 |
| 47 2aza[b] | 47.3 |
| 48 2cab | 74.6 |
| 49 2cro | 75.4 |
| 50 2ccy[a] | 82.7 |
| 51 2cdv | 73.8 |
| 52 2ci2[i] | 47.7 |
| 53 2cpp | 71.9 |
| 54 2cts | 71.6 |
| 55 2cyp | 69.3 |
| 56 2fb4[h] | 69.4 |
| 57 2gbp | 69.3 |
| 58 2gls[a] | 69.4 |
| 59 2gn5 | 67.8 |
| 60 2hhb[b] | 45.9 |
| 61 2ins[a] | 52.4 |
| 62 2ins[d] | 79.3 |
| 63 2lbp | 66.8 |
| 64 2lzm | 72.0 |
| 65 2ovo | 64.3 |
| 66 2pab[a] | 44.7 |
| 67 2plv[a] | 77.7 |
| 68 2sns | 66.0 |
| 69 2sod[b] | 78.1 |
| 70 2ssi | 71.0 |
| 71 2stv | 62.0 |
| 72 2taa | 75.1 |
| 73 2tbv[c] | 63.2 |
| 74 351c | 76.8 |
| 75 3adk | 72.7 |
| 76 3b5c | 64.7 |
| 77 3cla | 68.1 |
| 78 3cln | 85.3 |
| 79 3cna | 71.7 |
| 80 3cpv | 63.0 |
| 81 3fxc | 81.6 |
| 82 3fxn | 76.8 |
| 83 3gap[a] | 67.8 |
| 84 3grs | 64.2 |
| 85 3icb | 89.3 |
| 86 3icd | 72.9 |
| 87 3pgk | 72.0 |
| 88 3pgm | 79.1 |
| 89 3rp2[a] | 60.7 |
| 90 3rxn | 84.6 |
| 91 3tln | 67.4 |
| 92 3wrp | 74.3 |
| 93 3xia | 61.0 |
| 94 4cpa[i] | 73.0 |
| 95 4dfr[b] | 57.9 |
| 96 4fd1 | 67.9 |
| 97 4mdh[a] | 68.2 |
| 98 4pti | 67.2 |
| 99 4sbv[c] | 53.6 |
| 100 4ts1[a] | 71.9 |
| 101 5cpa | 67.8 |
| 102 5ldh | 59.2 |
| 103 6acn | 67.9 |
| 104 7api[a] | 59.6 |
| 105 7wga[b] | 81.2 |
| 106 8adh | 65.8 |
| 107 8atc[a] | 65.2 |
| 108 8atc[b] | 59.6 |
| 109 8cat[a] | 68.9 |
| 110 9pap | 69.3 |

## Table 6

Prediction accuracy for secondary structure predictors

| Method | Accuracy (%) |
|---|---|
| Bayes<br>(Stolarz *et al.*, 1992) | 61.1 |
| Homologue<br>(Levin *et al.*, 1986) | 62.2 |
| GOR III<br>(Gibrat *et al.*, 1987) | 63 |
| Neural Network<br>(Qian & Sejnowski, 1988) | 64.3 |
| Rooman-Wodak<br>(Rooman & Wodak, 1988) | 62 |
| King-Sternberg<br>(King & Sternberg, 1990) | 60 |
| GOR-COMBINED<br>(Biou *et al.*, 1988) | 65.5 |
| MBR-STM-NN-HYBRID<br>(Zhang *et al.*, 1992) | 66.4 |
| Nearest-Neighbor<br>(Yi & Lander, 1993) | 68.0 |

Accuracy as reported by the authors.

Table 7

Analysis of prediction accuracy at selected positions

| Subset of Predicted Sites | Accuracy (%) |
|---|---|
| A) | |
| 1. $\underline{\text{h}}$ | 62.6 |
| 2. $\underline{\text{e}}$ | 41.7 |
| 3. $\underline{\text{c}}$ | 80.8 |
| B) | |
| 1. h$\underline{\text{h}}$h | 68.6 |
| 2. e$\underline{\text{e}}$e | 49.9 |
| 3. c$\underline{\text{c}}$c | 82.5 |
| 4. c$\underline{\text{h}}$h or h$\underline{\text{h}}$c | 36.8 |
| 5. c$\underline{\text{e}}$e or e$\underline{\text{e}}$c | 30.2 |

Helix, β-strand, and coil are represented respectively by 'h', 'e', and 'c'. The predicted position is underlined. The known secondary structure at and adjacent to the predicted position is depicted.

## Table 8

Altering the radius of the exemplars in an exemplar-based nearest-neighbor system

| | Pruning Exemplars (% Removed) | Expanding Exemplars (% Expanded) | Accuracy (%) |
|---|---|---|---|
| 1 | 0 | 0 | 66.7 |
| 2 | 10 | 0 | 66.9 |
| 3 | 20 | 0 | 66.8 |
| 4 | 30 | 0 | 66.6 |
| 5 | 10 | 10 | 67.3 |
| 6 | 20 | 20 | 67.3 |
| 7 | 30 | 30 | 67.2 |
| 8 | Standard Nearest-Neighbor Method (k = 50) | | 66.8 |

The initial radius for all exemplars was 50. The scoring system was "Second/Access/Polar(15) + Benner." The window size was 13. The radius of the pruned exemplars was set to 0. The radius of the expanded exemplars was increased to 100.

Table 9

An analysis of the estimated probability values

| Probability Value of Prediction (%) | | Accuracy (%) | Proportion of Database (%) |
|---|---|---|---|
| A) | | | |
| > 80 | (87.8) | 85.9 | 28.4 |
| 70 - 80 | (74.8) | 72.6 | 14.9 |
| 60 - 70 | (64.9) | 64.6 | 25.7 |
| 50 - 60 | (55.1) | 55.7 | 16.6 |
| ≤ 50 | (45.6) | 43.9 | 14.5 |
| B) | | | |
| 40 - 50 | (44.8) | 43.4 | |
| 30 - 40 | (33.9) | 34.3 | |
| 20 - 30 | (23.7) | 23.8 | |
| 10 - 20 | (13.5) | 13.4 | |
| 0 - 10 | ( 4.7) | 5.7 | |

The probability values were placed into intervals spanning 10 percentage points. The number in parenthesis refers to the average value for that interval. The accuracy was calculated by counting the number of times that the secondary structure type associated with the estimated probability value was correct. In part A), the estimated probability values represent the highest value in the probability triplet at a given position. Proportion of the database refers to the fraction of the predictions in which the maximum estimated probability value at a given position fell within the specified interval. In part B), the probability values were not necessarily the maximum value in the triplet.

## Table 10

A breakdown of the probability values by prediction type

| Probability Value of Prediction (%) | Proportion of Each Type of Prediction (%) | | |
|---|---|---|---|
| | Helix | Beta-sheet | Coil |
| > 80 | 24.4 | 11.6 | 34.3 |
| 70 - 80 | 14.5 | 9.9 | 16.3 |
| 60 - 70 | 29.0 | 27.8 | 23.6 |
| 50 - 60 | 18.3 | 24.9 | 13.8 |
| ≤ 50 | 13.8 | 25.8 | 12.1 |

The probability values were divided according to the secondary structure type associated with the value. They were then placed into intervals spanning 10 percentage points. Finally, the proportion of each type of prediction that fell into the different probability intervals was measured.

## Acknowlegments

We thank Drs. Bruce Tidor, Xiru Zhang, Lydia Gregoret, and Rick Lathrop for critically reading the manuscript. We thank Drs. Bruce Tidor and Xiru Zhang for enlightening discussions.

# Chapter 4

# Recognition of Related Proteins by Iterative Template Refinement (ITR)

## 4.1 Abstract

Predicting the structural fold of a protein is an important and challenging problem. Available computer programs for determining whether a protein sequence is compatible with a known three-dimensional structure fall into two categories: (i) *structure-based methods*, in which structural features such as local conformation and solvent accessibility are encoded in a template, and (ii) *sequence-based methods*, in which aligned sequences of a set of related proteins are encoded in a template. In both cases, the programs use a static template based on a predetermined set of proteins. Here, we describe a computer-based method, called Iterative Template Refinement (ITR), that uses templates combining structure-based and sequence-based information and employs an iterative search procedure to detect related proteins and sequentially add them to the templates. Starting from a single protein of known structure, ITR performs sequential cycles of database search to construct an expanding tree of templates with the aim of identifying subtle relationships among proteins. Evaluat-

ing the performance of ITR on six proteins, we found that the method automatically identified a variety of subtle structure similarities to other proteins. For example, the method identified structural similarity between arabinose-binding protein and phosphofructokinase, a relationship that has not been widely recognized.

## 4.2 Introduction

The number of distinct protein structural folds is thought to be relatively small, perhaps less than 1000 (Chothia, 1992). If so, many of the 60,000 proteins in existing sequence databases must adopt conformations similar to an already known structure. However, it remains a challenging problem to place proteins known only by their amino acid sequences into the correct structural family.

Various computer programs have been developed for using a fixed template for detecting structurally related proteins. Broadly speaking, they fall into two categories.

(i) *Structure-based methods.* In this approach, a template encoding information about a specific structure is used to search the protein database to find other sequences compatible with the structure. The first such approach involved classifying each position in a structure according to its solvent accessibility, local conformation, and polarity (Bowie et al., 1991). Subsequent generalizations have included other types of environments defined according to the nature of contacts made by each residue (Ouzounis et al., 1993) as well as residue-residue contact potentials (Grodzik et al., 1992; Jones et al., 1992). These methods have been recently reviewed (Bowie & Eisenberg, 1993; Wodak & Rooman, 1993).

(ii) *Sequence-based methods.* Sequence similarity can often be a reliable indicator of structural homology (Chothia & Lesk, 1986). To detect very high degrees of sequence similarity, rapid pairwise comparison programs such as FASTA (Pearson & Lipman, 1988) and BLAST (Altschul et al., 1990) can be used. To detect more subtle similarities, many investigators have adopted the strategy of constructing a template based on the sequences of multiple members of a known protein family in order to focus attention on the most conserved features – thereby extracting signal from noise (e.g., Taylor, 1986; Bashford et al., 1987; Gribskov et al., 1987; Altschul & Lipman, 1990; Barton & Sternberg, 1990; Henikoff & Henikoff, 1991).

In contrast to such computer programs, experienced protein researchers searching for new members of a structural family adopt a broader approach. Investigators will typically use both structure-based and sequence-based information and will often examine the information in an iterative fashion – e.g., using comparisons of close family members to identify important structural or sequence features that permit subsequent recoginition of more distant family members.

Here, we describe a computer-based method that attempts to incorporate the ideas of (i) integration of structure and sequence information and (ii) iterated search. Called Iterative Template Refinement (ITR), the procedure starts with a single protein of known structure and derives a 'Level 1' template reflecting information about both structure and sequence of the protein. The Level 1 template is used to search the database to identify similar proteins, which are then used to construct various different Level 2 templates. The process is iterated to yield an expanding tree of dynamically-refined templates. Because template construction is dynamic, the method offers the prospect of discovering new similarities beyond the usual family groupings. Because the method is fully automated, user involvement is kept to a minimum.

To evaluate this approach, ITR was applied to six starting proteins: (i) arabinose-binding protein (1ABP); (ii) plastocyanin (1PCY); (iii) cytochrome c (1CCR); (iv) chymotrypsin (2CGA); (v) the dinucleotide-binding domain of lactate dehydrogenase (5LDH, domain 1); and (vi) the $\alpha$-subunit of tryptophan synthase (1WSY_A). In each case, ITR was able to detect similarities to structurally related proteins that could not be found by standard single sequence comparison. For example, templates constructed from plastocyanin automatically identified the structurally similar immunoglobulins and templates derived from bovine chymotrypsin automatically identified a variety of bacterial serine proteases. A particularly striking example was the finding of structural similarity between arabinose-binding protein and phosphofructokinase, a similarity that has not been widely recognized despite the availability of the two structures. Overall, ITR provides a potentially powerful approach for detecting subtle, structurally important similarities among proteins.

## 4.3 Outline of Methodology

ITR essentially consists of repeated cycles of database search and new template generation. Briefly, one starts with a Level 1 template based on a single protein of known structure. One searches the database with this Level 1 template to identify and align all significant matches in the protein sequence database. Each of the matches is used to derive a distinct Level 2 template, which are then used to search the database once again. The process is repeated through as many levels as feasible (Figure 1A), with each Level $k$ template potentially giving rise to *multiple* distinct Level $k+1$ templates. ITR thus yields an expanding tree of templates (Figure 1B). It should be noted that only the starting protein is required to have a known structure.

To make matters precise, let $P_1$ be a protein of known structure. In the manner of Bowie et al. (1991), each residue of $P_1$ is assigned an "environment class" describing its local neighborhood in the structure. The specific assignment used here is taken from Yi and Lander (1993), which uses 15 environment classes. A **Level k template of length m rooted at $\mathbf{P_1}$** is defined as a multiple subsequence alignment of the environment string and the amino acid sequence of $P_1$, together with the amino acid sequences of $k-1$ other proteins, $P_2, ..., P_k$. A Level 1 template thus involves just the environment string and the amino acid sequence of $P_1$. A Level 4 template of length 80 rooted at arabinose-binding protein (1ABP) is shown in Figure 1C.

A **multiple subsequence alignment** is formally defined by a $k \times m$ matrix $S$ whose *i-th* row, $(s_{i1}, ..., s_{im})$, is a sequence of consecutive positions in protein $P_i$, possibly interrupted by occurrences of the gap symbol "-". Let $a_{ij}$ denote the amino acid in position $s_{ij}$ in protein $P_i$ or the gap symbol if $s_{ij}$ is a gap symbol. Similarly, let $e_j$ denote the environment symbol of the amino acid in position $s_{1j}$ in protein $P_1$ or the gap symbol if $s_{1j}$ is a gap symbol.

ITR involves three steps, with the first two applied at each successive Level $k$.

(i) *Database search.* To search the protein sequence database with a Level $k$ template $T$, the template is converted into a **profile** using an extension of the method of Gribskov et al. (1987). The profile combines the sequence and structure information of the template into a single scoring function, $PROFILE_j(x)$, that specifies the score for matching the *j-th* position in the template to amino acid $x$ in a target protein $P$. Specifically, the profile is defined as:

$$PROFILE_j(x) = \alpha \sum_{i=1}^{k} A(a_{ij}, x) + \beta B(e_j, x), \tag{4.1}$$

where $A(\bullet, \bullet)$ is a traditional amino acid by amino acid scoring matrix and $B(\bullet, \bullet)$ is an amino acid by environment scoring matrix (see Materials and Methods). Thus, the profile makes possible the alignment of a target sequence against a template containing both sequence and structure information. The parameters $(\alpha, \beta)$ control the relative weight to be placed on sequence versus structure; they were chosen so as to place twice as much weight on the sequence component as on structure (see Materials and Methods). A traditional Gribskov profile would correspond to setting $\beta = 0$.

The resulting profile is compared to each protein P in the database by using a standard 'local' dynamic programming alignment algorithm (Smith & Waterman, 1981) with appropriate gap penalties, $g_{open}$ and $g_{ext}$, for opening and extending a gap. For each protein P, the resulting score $Y = Y(T, P)$ is converted into a 'normalized' Z-score by using a standard Monte Carlo shuffling test (Doolittle, 1986; Gribskov & Devereaux, 1991; Karlin et al., 1991). Proteins $P$ with final Z-scores exceeding 7.5 are considered to be significant matches to the template $T$. (See Materials and Methods concerning choice of parameters and thresholds).

(ii) *New template generation.* Significant matches resulting from a search with a Level $k$ template are used to derive distinct Level $k + 1$ templates, which are then used for a subsequent round of database search.

(iii) *Analysis of completed search tree.* An ITR search produces a tree of templates, which can be represented as a graph (Figure 1B). Using this graph, the final 'significance score' for each protein found in the ITR search is defined as follows. Each edge connecting a Level $k$ template to a Level $k + 1$ template is labelled with the Z-score for the protein added in creating the Level $k + 1$ template. Each node of the tree is labelled with the smallest (i.e., least significant) Z-score on the path connecting the node to the root of the tree. A particular protein $P$ may occur at multiple nodes in the tree; the final 'significance score' for $P$ is defined as the maximum of the labels for nodes at which $P$ occurred.

As outlined above, ITR is conceptually simple but computationally inefficient. The reasons are several: (a) the search tree can expand exponentially at each level, particularly as many related proteins are identified; (b) search of the entire database using dynamic programming is very time-consuming; and (c) complete Monte Carlo shuffling to calculate a Z-score for each match is too slow. To speed up the procedure, we adopted several computational compromises outlined in Material and Methods.

## 4.4 Results

To assess the accuracy and sensitivity of ITR, we tested the approach using six starting proteins whose structures are known: (a) arabinose-binding protein (1ABP); (b) plastocyanin (1PCY); (c) cytochrome c (1CCR); (d) chymotrypsin (2CGA); (e) lactate dehydrogenase, Rossman domain (5LDH, domain 1); and (f) tryptophan synthase, $\alpha$-subunit (1WSY_A). We examined the list of matching proteins identified by ITR. For proteins with known structures, the accuracy of the match could be rigorously assessed by comparing the structures of the identified protein and the starting protein. The structures were aligned by using a computer program (T. -M. Yi, unpublished) to minimize DRMSD and were considered structurally related if the alpha carbon atoms could be aligned at $\geq 60\%$ of the positions in the smaller protein with Distance Matrix RMSD (DRMSD) of $\leq 2.7$ Å. For proteins without known structures, the accuracy of the match could only be assessed qualitatively (and thus somewhat subjectively) based on information about the structure, function, and regulation of the protein.

### (a) Arabinose-binding protein (1ABP)

Arabinose-binding protein is a two domain protein belonging to the family of periplasmic binding proteins that includes galactose-binding protein, ribose-binding protein, phosphate-binding protein, etc. (Spurlino et al., 1991). The structure of arabinose-binding protein (1ABP) was solved to a resolution of 2.4 Å by Gilliland and Quiocho (1981).

The Level 1 template identified only two significant matches: ribose-binding protein and galactose-binding protein (Figure 2A). Each was separately aligned to the Level 1 template to create two Level 2 templates. Using these Level 2 templates, a new class of proteins emerged: members of the lac repressor family. For example, the Level 2 template *1ABP_JGECR*, produced nine significant matches (see Figure 2B). As the search progressed, the branching of the search tree was constrained to prevent an exponential explosion of templates (see Materials and Methods). A typical Level

4 template is shown in Figure 2C. The complete search results are shown in Table 1.

One interesting match was to the periplasmic leucine-binding protein (2LBP), for which the structure is available. Despite low sequence similarity, structural alignment confirmed that leucine-binding protein and arabinose-binding protein are structurally homologous (DRMSD = 2.4 Å, aligning 214 alpha carbon atoms).

A more surprising match was to two families of transcriptional regulators: the lac repressor family and the two-component response regulator family. The similarity with the former family had been noted by Muller-Hill (1983) based on careful manual sequence analysis of these proteins and is supported by the fact that both classes of proteins bind small ligands (often a sugar moiety) and undergo a conformational change upon binding. Using ITR, this similarity emerged automatically. To our knowledge, the similarity to the two-component response transcriptional regulators has not been previously noted. Interestingly, these response proteins also bind a small ligand (ATP) and trigger the conformational switch of the transcriptional apparatus from closed to open complex (Kustu et al., 1991). Because structures are not available for these families, these structural similarities cannot be confirmed.

The most unexpected match was to phosphofructokinase, identified on level 4. Although the structures of arabinose-binding protein and phosphofructokinase have been known for years, similarity between them has not been widely noted. In fact, the two proteins are structurally quite similar (DRMSD = 2.4 Å, aligning 184 alpha carbon atoms). Functional and structural clues suggest a potential relationship. Both proteins are composed of two $\alpha/\beta$ domains, and both proteins bind their ligand(s) in the cleft between the two domains. Upon binding, both proteins undergo a conformational shift in which the two domains pivot to enclose the ligand. (We are grateful to a referee for pointing out that Holm and Sander (1993) have recently performed a hierarchical clustering of all known structures, in which the periplasmic binding proteins and phosphofructokinase appear in the same region of the tree.)

In summary, the search identified three proteins with known structures. All three matches were confirmed by structural alignment. In two cases (galactose-binding protein and leucine-binding protein), the similarity was strong enough that it could be found by standard pairwise sequence comparison or the Bowie-Eisenberg method, while in the third case (phosphofructokinase) the similarity was too subtle to be recognized by such pairwise comparisons.

## (b) Plastocyanin (1PCY)

Plastocyanin is a small copper binding protein that plays a role in electron transfer. The Level 1 template, derived from poplar plastocyanin (1PCY), identified plastocyanins from other species and the closely-related pseudoazurins and amicyanins (Table 2). The Level 3 templates identified two other classes of proteins involved in electron transport: azurins and basic blue proteins. Although these proteins have low sequence similarity to plastocyanin, their structures have been solved and are known to be very similar to 1PCY (Baker, 1988; Guss et al., 1988). Interestingly, the Level 3 templates also uncovered a class of proteins functionally far removed from plastocyanin: various members of the immunoglobulin superfamily (see Table 2). In fact, the structure of plastocyanin has been noted to resemble the $\beta$-barrel topology of the "immunoglobulin fold" (Guss & Freeman, 1983). Finally, the Level 4 templates identified a significant match with ascorbate oxidase, a blue multi-copper oxidase. In fact, the structure ascorbate oxidase was recently solved by Messerschmidt et al. (1992), and it possesses three domains of a similar $\beta$-barrel type as plastocyanin (DRMSD = 2.0 Å, aligning 94 alpha carbon atoms).

Of the identified proteins, four groups possessed members with known structure (azurins, basic blue protein, immunoglobulins, and ascorbate oxidase). The accuracy of these matches was confirmed in all four cases. Of the four, only the similarity to the immunoglobulins could be detected using standard pairwise comparison methods.

## (c) Cytochrome c (1CCR)

The c-type cytochromes have been grouped into four classes based on structural and electrochemical criteria (Ambler, 1991). The members of each class are structurally homologous to one another, while the members of different classes possess different structural topologies. Class I, which includes the main group of mitochondrial cytochrome c proteins, has been further subdivided into five subclasses according to sequence similarity. A Level 1 template was derived from rice cytochrome c (1CCR), a member of Class I, subgroup B. The results are shown in Table 3. The initial search identified other members of subgroups A and B (primarily cytochrome c, c2, and c550 proteins). Subsequent levels identified members of subgroups C (cytochrome c6), D (cytochrome c551), and E (cytochrome c555). Most of these proteins have low sequence similarity with rice cytochrome c, and historically, structural data was necessary to confirm the close relationship among the different class I subgroups (Ambler, 1991). The search also identified the protein p-cresol methylhydroxylase which is known to consist of two subunits: a flavoprotein and a c-type cytochrome (McIntire et al., 1986). Analysis of the crystal structure of this protein has verified that the cytochrome subunit belongs to the class I family of cytochrome c proteins (Mathews et al., 1991).

Of the identified proteins, three have known structures. The accuracy of the matches was confirmed in all cases. None of these could be detected using standard pairwise comparison methods.

## (d) Chymotrypsin (2CGA)

A Level 1 template, derived from bovine chymotrypsin (2CGA), not surprisingly identified many of the members of the mammalian serine protease family. As the search proceeded, several different types of bacterial proteases emerged – including members of both the streptomyces and staphylococcus classes of serine proteases, along with *B. subtilis* metalloproteinase, *Achromobacter* proteinase I, and heat shock protein HtrA (Table 4). In the past, it was suspected from functional data that some of the bacterial proteases were structurally related to the mammalian serine proteases, but sequence analysis proved inconclusive (Delbaere et al., 1979). For the streptomyces family of serine protease, the question was resolved when several structures were solved ($\alpha$-lytic protease, proteases A and B from *S. griseus*) and shown to be quite similar to the structures of the mammalian serine proteases (Fujinaga et al., 1985). Moreover, the structure of *Achromobacter* proteinase I has been solved recently, and it is quite similar to chymotrypsin (DRMSD = 2.1 Å, aligning 197 alpha carbon atoms). Here, ITR was able to rediscover and potentially extend the relationship between bacterial and mammalian serine proteases.

Of the identified proteins, two have known structures. The accuracy of the matches was confirmed in both cases. Neither could be detected using standard pairwise comparison methods.

## (e) Lactate dehydrogenase – dinucleotide-binding domain (5LDH, domain 1)

The Rossmann fold is a common structural topology, consisting of a parallel $\beta$-sheet with $\alpha$-helices packed on both sides. It is found in many proteins that bind NAD, NADP, ATP, GTP, and FAD (Rossmann, 1974). A search was initiated with the Rossman domain of lactate dehydrogenase from dogfish (residues 22 to 164). The Level 1 template identified lactate dehydrogenases from other species along with the closely-related malate dehydrogenases. At Level 2, several other dehydrogenases began to appear along with a variety of GTP-binding proteins. By the later stages of the search, a large number of different dehydrogenases, reductases, and hydrolases were detected. In all, 15 classes of proteins were identified, most of which are known to bind NAD, NADP, ATP, or GTP (Table 5).

Structural data is available for four of the identified proteins: phosphoglycerate kinase, G3P dehydrogenase, 3-hydroxyacyl-CoA dehydrogenase and p21 ras protein. The first three proteins clearly possess the Rossmann fold, while the structure of p21 ras diverges somewhat from the canonical topology. Nonetheless, it is possible to align 97 alpha carbon atoms (out of 143 residues) between the two structures to 2.5 Å DRMSD, which satisfies the criterion for structural similarity stated above.

### (f) Tryptophan synthase – alpha subunit (1WSY_A)

The tryptophan synthase $\alpha$-subunit catalyzes the last reaction in tryptophan biosynthesis. It has a canonical TIM barrel structure, consisting of a $\beta$-sheet wrapped into a barrel surrounded by $\alpha$-helices. A Level 1 template derived from tryptophan synthase $\alpha$-subunit from Salmonella typhimurium (1WSY_A) identified only other bacterial tryptophan synthases.

Interestingly, Level 4 templates identified another enzyme in the tryptophan biosynthesis pathway, indole-3-glycerol-phosphate synthase (TrpC), and two enzymes involved in histidine biosynthesis, cyclase HisF and compound III isomerase (HisA) (see Table 6). The structure of indole-3-glycerol-phosphate synthase is known, and it is indeed a TIM barrel protein. Work by Wilmanns and Eisenberg (1993) using a more advanced version of the Bowie-Eisenberg 3D profile method have recently suggested that cyclase HisF and compound III isomerase also possess a TIM barrel structure.

Various other proteins emerged including several oxidases that use NAD as a cofactor ((S)-2-hydroxy-acid oxidase, flavocytochrome B2, and dihydroorotate oxidase) and several ion transporting ATPases ($H^+/K^+$-transporting ATPase, $Ca^{2+}$-transporting ATPase, and cadmium-transporting ATPase). The structures of (S)-2-hydroxy-acid oxidase (DRMSD = 2.0 Å, aligning 194 alpha carbon atoms) and flavocytochrome B2 (DRMSD = 2.5 Å, aligning 162 alpha carbon atoms) are known to be TIM barrels; no structural data are available for the ion channel proteins.

Finally, the search produced the first two definite false positives. Level 4 templates identified two $\alpha/\beta$ proteins, arabinose-binding protein and p21 ras protein, with Z-scores just above the threshold of 7.5. Although both proteins possess an alternating $\alpha$-helix/$\beta$-strand secondary structure pattern such as is found in tryptophan synthase, the overall tertiary folds are different. We discuss below the reasons for these false positives. Thus, among the known structures, this search recorded two correct hits and two false positives.

### (g) Searching based on sequence alone

In principle, ITR could be performed without making use of the structure-environment component (i.e., setting the weighting parameter $\beta = 0$). To test whether the structure-environment component contributed significantly to the sensitivity of the method, we repeated the searches for arabinose-binding protein and plastocyanin without this information. The results are presented in the last column of Tables 1 and 2. It is clear that there is a substantial loss of sensitivity.

## 4.5 Discussion

Iterative Template Refinement is intended to be an automated procedure for identifying proteins distantly related to a starting protein. By dynamically creating a tree of templates, ITR takes a bootstrapping approach to extracting key features in a protein – letting them emerge spontaneously as templates are refined through successive database searches. In this way, ITR can find matches that are not detected by traditional single-protein comparison techniques. Of course, the crucial question is: Do the reported matches actually represent structural similarity among distantly related proteins?

### (i) Accuracy

The best test of accuracy is to examine matches to proteins of known structure to evaluate whether there is true structural similarity. In the six searches reported above, there were 20 such matches (Table 7). Of these, the identified protein had clear structural similarity in 18 cases and was unrelated in two cases. We note that our definition of structural similarity (DRMSD $\leq$ 2.7 Å aligning at least 60% of the alpha carbon positions) is somewhat less strict than that employed by some crystallographers, but is designed to capture the notion that two proteins possess the 'same' structural fold. Indeed, this criterion reflects the level of structural similarity between members of large structural families such as the jellyroll $\beta$-barrel family (Chelvanayagam et al., 1992) and the Greek key $\beta$-barrel family (Hazes & Hol, 1992). Among the 18 confirmed matches, only three were also detected by FASTA, BLAST, or the Bowie-Eisenberg method (Bowie et al., 1991). Thus, ITR was able automatically to discover various subtle relationships, with a reasonably low (but non-zero) false positive rate.

The two false positives occurred in the search with tryptophan synthase $\alpha$-subunit and had Z-scores just above the threshold for significance (arabinose-binding protein and p21 ras, each with $Z = 7.5$). In examining the three structures, we noted that they do share some charateristics: they are all $\alpha/\beta$ proteins consisting of repeated units of $\alpha$-helix and $\beta$-strand and, in fact, they can be aligned in a manner consistent with these similarities in secondary structure, solvent accessibility, and even sequence (not shown). Nonetheless, the overall topologies are different. This observation suggests that ITR might be further improved by including a final evaluation step based on three-dimensional contact potentials, which encode distance relationships between different positions in a structure (Jones et al., 1992).

Importantly, ITR misses many structural similarities among proteins. Indeed, there remains no automatic way to recognize all distant similarities among proteins. For example, the arabinose-binding protein search failed to identify many other periplasmic binding proteins, which are known to have structural similarity (Spurlino et al., 1991). The plastocyanin search identified the immunoglobulins, but missed many other proteins with the immunoglobulin fold such as superoxide dismutase and actinoxanthin. Similarly, the tryptophan synthase search did not recognize many TIM barrel proteins including TIM itself, enolase, and $\alpha$-amylase. At this early stage, however, we were more concerned with minimizing false positives than false negatives; the choice of a relatively high threshold cutoff of 7.5 reflects this concern.

For most of the matches, the accuracy cannot be rigorously evaluated because no structural information is available for the target protein. These matches should be regarded as a test set that will permit unbiased evaluation of the method as the structures are eventually solved. In the meanwhile, they offer intriguing hypotheses that may provoke further investigations. For example, the searches reported above suggest three interesting predictions: (i) the two-component transcriptional regulatory proteins are structurally related to the lac repressor farmily; (ii) the staphylococcus family of bacterial proteases possess the same structural fold as the mammalian serine proteases; and (iii) the $Na^+/K^+$-ion-transport ATPases contain a TIM barrel domain.

## (ii) Relationship with other methods

ITR combines both structure-based and sequence-based approaches to the detection of similarities and uses iteration to automatically derive templates without prior specification of a protein family. Clearly, ITR draws extensively on prior work by many investigators (Taylor, 1986; Gribskov et al., 1987; Altschul & Lipman, 1990; Barton & Sternberg, 1990; Smith & Smith, 1990; Bowie et al., 1991; Henikoff & Henikoff, 1991).

ITR differs from most previous structure-based approaches in that it makes extensive, iterative use of sequence information. A single static template combining both sequence and structure information was previously used by Pickett et al. (1992) in a study of TIM barrel proteins, although they concluded that the combination did not improve discrimination over the use of sequence information alone. It should be noted that the study involved only a single hybrid template; there was no iteration.

ITR differs from most previous sequence-based approaches in that templates are automatically derived without prior specification of a protein family. Iterative template construction was first explored by Taylor (1986) in his study of immunoglobulins. However, in this and similar studies, the proteins studied were all identified a priori. Because ITR involves construction of an expanding tree of templates, it can potentially identify new relationships.

Altschul and Lipman (1990) were among the first to recognize the importance of using multiple sequence alignments as a database search tool, noting that it is a strategy used implicitly by experienced protein researchers. They developed a rapid heuristic search program BLAST3 implementing database search based on 3-way alignments. ITR is conceived in a similar spirit, but takes a rather different approach by using a multi-level expanding tree search, which may successively refine features in a template.

It should be noted that both the structure-based component and the sequence-based iteration play an important role in ITR. Omitting the structure-based information greatly weakens the sensitivity of the method, as noted in the Results section. Similarly, failing to search in an iterative fashion (i.e., using only Level 1 searches) also greatly diminishes the number of matches detected.

## (iii) Limitations and future directions

There are many open questions about the optimal implementation, sensitivity, and selectivity of this search method. The current program is still quite slow, requiring one week per protein on a standard workstation and we are investigating algorithmic improvements and heuristic shortcuts to speed up the search without unacceptable loss of sensitivity. In addition, the program involves a variety of parameter choices for scoring matrices, gap penalties, etc. The choices are reasonable, but have not been optimized.

More importantly, ITR produces some clear false positives and many false negatives. Concerning false positives, it will be important to study the various parameters and thresholds used in the procedure. The current definition of a "significant" match remains empirical and somewhat ad hoc; a useful statistical theory for the significance of tree search is currently lacking and is likely to be challenging. Concerning false negatives, it may be possible to add additional structural information such as a contact pair potential or predictions of secondary structure. Indeed, many of the recent advances in structure-based methods to inverted prediction could be adapted to ITR (Goldstein et al., 1992; Grodzik et al., 1992; Jones et al., 1992; Sippl et al., 1992; Ouzounis et al., 1993). More broadly, it will be important to test ITR on a more extensive list of examples to evaluate its performance. At present, ITR should be viewed as a tool for exploratory data analysis producing putative matches requiring confirmation.

## 4.6 Materials and Methods

### (A) Compromises required for efficient computation

As outlined in the text, ITR is conceptually simple but computationally inefficient. The reasons are several: (a) the search tree can expand exponentially at each level, particularly as many related proteins are identified; (b) search of the entire database using dynamic programming is very time-consuming; and (c) complete Monte Carlo shuffling to calculate a Z-score for each match is too slow. To speed up the procedure, several computational compromises were adopted. Many alternative choices could clearly be made, but the chosen shortcuts seemed to provided adequate speed up without noticeable loss of sensitivity. However, these choices cannot be said to be optimal in any sense.

(i) *Pruning the search tree.* To prevent the tree from branching too extensively, three constraints were used for pruning. (a) The search was terminated after the construction of Level 6 templates. (b) The total number of Level $k$ templates used to search the database was limited to 6 on Level 2; 30 on Level 3; and 48 on Levels 4 and 5. If the available number of Level $k$ templates exceeded this bound, only the 'most promising' templates were used. Templates were ranked based on the total number of matching proteins having significant Z-scores and the total number of matching proteins having Z-scores that had increased from a previous level. (c) The total number of Level $k+1$ templates generated from any given Level $k$ template was limited to 6 on Level 1; 5 on Level 2; 4 on Level 3; and 3 on Level 4. If the number of significant database matches for a given template exceeded this bound, a modified version of the hierarchical clustering algorithm of Smith & Smith (1990) was used to group the proteins into the maximum allowed number of clusters. Briefly, the clustering algorithm involves successively grouping together pairs of sequences with the highest similarity. One representative from each cluster was chosen. This eliminated the explosion caused by related members of protein families and ensured that the paths chosen were as different as possible.

(ii) *Pre-screening the database.* Because dynamic programming is very time-consuming, templates were initially compared to the entire protein database by using a modified version of the BLAST algorithm (sBLAST) to identify the highest-scoring 2000 proteins. For a given template, we calculated a list of the highest scoring 4-mers (Altschul et al., 1990), scanned the database for proteins containing words in the list, and extended the hits into complete matches. The highest-scoring 2000 proteins identified in this manner were subsequently analyzed by complete dynamic programming.

(iii) *Calculating Z-scores.* When a template T was compared to a protein P of length n, the resulting score $Y = Y(T, P)$ was converted to a Z-score by a standard Monte Carlo shuffling technique (Doolittle, 1986; Karlin et al., 1991). The template $T$ was compared to a collection of 600 random sequences having the same length as $P$, with amino acids randomly chosen according to the frequency distribution in the database. The mean yn and standard deviation $\sigma_n$ for these random comparisons were determined and the Z-score was defined as $Z = \frac{(Y - y_n)}{\sigma_n}$. To avoid calculating a Z-score for every possible length $n$, we first calculated the mean $y_n$ and standard deviation sn for sequences of length n = 100, 200, 400, 600, and 800. The results for other values of n were estimated by interpolation and the Z-score was estimated using these interpolated values. If the estimate exceeded 6.5, the Z-score was then calculated using Monte Carlo shuffling for the exact length.

The overall distribution of Z-scores was bimodal, with a large peak extending from $Z = -3.0$ to $Z = 3.0$ consisting primarily of unrelated sequences, and a second smaller peak for Z-scores above 12.0 consisting primarily of very closely related sequences. The range of Z-scores between $Z = 6.0$ to $Z = 9.0$ appeared to contain many interesting matches to potential distant relatives and the threshold of significance was chosen to balance between the false positives and false negatives.

Matching proteins were tested for compositional bias by comparing the distribution of amino acids in the protein with the overall distribution in the database, using the chi-squared statistic as a measure of deviation. Only one protein showed an unusually large deviation: 125K hypothetical protein detected in the 1ABP search. When the specific sequence of this protein was used in the random permutation test, the resulting Z-score still exceeded the threshold ($Z = 8.5$). The current implementation of ITR permits calculation of Z-scores using the amino acid composition of either the target protein or the average composition of the database. In most cases, the two scores are quite similar (data not shown).

With these compromises, a complete ITR search for a single starting protein required about one week on a DEC Alpha workstation using the PIR sequence database, release 26.0. The execution time of ITR was approximately equally divided between searching the database and calculating the Z-scores. This time could be shortened by making further computational shortcuts. However, the performance was adequate for our present purposes of studying the method. The computer program implementing ITR as described above is available from the authors.

## (B) Parameter selection

(i) *Implementation of dynamic programming.* The local optimal dynamic programming algorithm of Smith-Waterman (Smith & Waterman, 1981) was implemented with two minor modifications. To avoid very short alignments, significant matches were required to span at least two-thirds of residues in the starting protein. Because the multiple sequence profile portion of the search template could contain gaps, the "pay once" gap penalty strategy described by Smith and Smith (1990) was used.

(ii)*Scoring matrices.* In equation (1), the amino acid by amino acid scoring matrix $A$ was chosen to be the Benner mutational distance matrix (Gonnet et al., 1992), and the amino acid by environment scoring matrix $B$ (Figure 3) was defined as in Yi and Lander (1993), although other choices are clearly possible.

(iii) *Weighting parameters.* The weighting parameters $(\alpha, \beta)$ were defined in terms of the 'magnitude' of the sequence versus structure-environment of each template. With the notation as in equation (1), the magnitude $m_A$ of the sequence component and $m_B$ of the structure-environment was defined as

$$m_A = \frac{1}{L} \cdot \sum_{j=1}^{L} \sum_{i=1}^{k} \sum_{x} max(A(a_{ij}, x), 0) \qquad (4.2)$$

and

$$m_B = \frac{1}{L} \cdot \sum_{j=1}^{L} \sum_{x} max(B(e_j, x), 0), \tag{4.3}$$

where the inner summations are taken over all amino acids x and L is the length of the template. For each template, the weighting parameters were then taken as $(\alpha, \beta) = (\frac{2}{3m_A}, \frac{1}{3m_B})$ – which has the effect of placing twice as much weight on the sequence component as on the structure component.

(iii) Gap penalties. Two sets of values for the gap penalties were used, $(g_{open}, g_{ext}) = (0.55, 0.11)$ and $(0.75, 0.15)$. A target sequence was aligned with a template using both gap penalties and the higher Z-score recorded. By way of comparison, normalizing the gap penalties $(20.63, 1.65)$ recommended by Gonnet et al. (1992) by dividing by the magnitude of their scoring matrix produces the values $(1.91, 0.15)$. Thus, the penalty for opening a gap is somewhat smaller than that used by Gonnet et al. (1992).

The choice of parameters was altered slightly after completing the searches with 1ABP, 1PCY, 1CCR. We had initially included an additional set of values for the weighting parameters $(\alpha, \beta)$, reflecting equal contributions of sequence and structure information, and for the gap penalties, $(g_{open}, g_{ext}) = (1.00, 0.20)$. These additional choices were dropped because they yielded poor alignments and poorer sensitivity. In addition, we slightly raised the threshold for significance from our initial choice of 7.0 to 7.5. The first three searches were rerun with these final parameters, with the result that two false positives were eliminated from the 1ABP search and one false positive from the 1CCR search. The latter three searches were performed only with the final parameters.

## 4.7 Appendix 1: Updating the List of Proteins of Known Structure Identified by ITR

Since the publication of this chapter in the journal *Protein Science* (Yi and Lander, 1994), two more proteins detected by the ITR searches have had their structures solved. In both cases, the structural data confirmed the prediction of structural similarity between the target protein(s) and the root protein. First, the structure of PurR, a member of the Lac repressor family of DNA-binding proteins, was determined to a resolution of 2.7 Å (Schumacher et al., 1994). The corepressor binding domain (CBD) or PurR displayed remarkable structural homology with the periplasmic binding proteins (the RMSD between PurR CBD and ribose-binding protein is 2.3 Å over 144 alpha carbons). Thus, as predicted in Table 1, the members of the Lac repressor family do indeed adopt the same fold as the root protein of the search, arabinose-binding protein. Secondly, Holm et al. (1994) noticed the structural relationship between $3\alpha$,$20\beta$-hydroxysteroid dehydrogenase (1HSD), a member of the short-chain dehydrogenase family (row number 3 in Table 5), and both dihydropteridine reductase (1DHR) and the classical dinucleotide-binding topology found in lactate dehydrogenase (5LDH). The common core between 1HSD and 5LDH extends for about 130 residues. These data strongly suggest that the short-chain dehydrogenase family and the Rossmann family share a common fold. Thus, among 22 matches to proteins of known structure, ITR registered 20 correct hits and two false positives.

# References

[Altschul et al., 1990] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.*, 215:403–410.

[Altschul and Lipman, 1990] Altschul, S. F. and Lipman, D. J. (1990). Protein database searches for multiple alignments. *Proc. Natl. Acad. Sci. U.S.A.*, 87:5509–5513.

[Ambler, 1991] Ambler, R. P. (1991). Sequence variability in bacterial cytochrome c's. *Biochim. Biophys. Acta*, 1058:42–47.

[Baker, 1988] Baker, E. N. (1988). Structure of azurin from *alcaligenes denitrificans.* Refinement at 1.8 Å resolution and comparison of the two crystallographically independent molecules. *J. Mol. Biol.*, 203:1071–1095.

[Barton and Sternberg, 1990] Barton, G. J. and Sternberg, M. J. E. (1990). Flexible protein sequence patterns: A sensitive method to detect weak structural similarities. *J. Mol. Biol.*, 212:389–402.

[Bashford et al., 1987] Bashford, D., Chothia, C., and Lesk, A. M. (1987). Determinants of a protein fold. Unique features of the globin amino acid sequences. *J. Mol. Biol.*, 196:199–216.

[Bowie and Eisenberg, 1993] Bowie, J. U. and Eisenberg, D. (1993). Inverted protein structure prediction. *Curr. Opin. Struct. Biol.*, 3:437–444.

[Bowie et al., 1991] Bowie, J. U., Luthy, R., and Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, 253:164–170.

[Chelvanayagam et al., 1992] Chelvanayagam, G., Heringa, J., and Argos, P. (1992). Anatomy and evolution of proteins displaying the viral capsid jellyroll topology. *J. Mol. Biol.*, 228:220–242.

[Chothia, 1992] Chothia, C. (1992). One thousand families for the molecular biologist. *Nature*, 357:543–544.

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

[Delbaere et al., 1979] Delbaere, L. T. J., Brayer, G. D., and James, M. N. G. (1979). Comparison of the predicted model of $\alpha$-lytic protease with the x-ray structure. *Nature*, 279:165–168.

[Doolittle, 1986] Doolittle, R. F. (1986). *Of urfs and orfs: a primer on how to analyze derived amino acid sequences.* University Science Books, Mill Valley, CA.

[Fujinaga et al., 1985] Fujinaga, M., Delbaere, L. T. J., Brayer, G. D., and James, M. N. G. (1985). Refined structure of $\alpha$-lytic protease at 1.7 Å resolution. Analysis of hydrogen bonding and solvent structure. *J. Mol. Biol.*, 184:479–502.

[Gilliland and Quiocho, 1981] Gilliland, G. L. and Quiocho, F. A. (1981). Structure of the L-arabinose-binding protein from *escherichia coli* at 2.4 Å resolution. *J. Mol. Biol.*, 146:341.

[Goldstein et al., 1992] Goldstein, R. A., Luthey-Schulten, Z. A., and Wolynes, P. G. (1992). Protein tertiary structure recognition using associative memory Hamiltonians with local interactions. *Proc. Natl. Acad. Sci. U.S.A.*, 89:9029–9033.

[Gonnet et al., 1992] Gonnet, G. H., Cohen, M. A., and Benner, S. A. (1992). Exhaustive matching of the entire protein sequence database. *Science*, 256:1443–1445.

[Gribskov and Devereux, 1991] Gribskov, M. and Devereux, J., editors (1991). *Sequence Analysis Primer.* Stockton Press, New York.

[Gribskov et al., 1987] Gribskov, M., McLachlan, A. D., and Eisenberg, D. (1987). Profile analysis: Detection of distantly related proteins. *Proc. Natl. Acad. Sci. U.S.A.*, 84:4355–4358.

[Grodzik et al., 1992] Grodzik, A., Kolinski, A., and Skolnick, J. (1992). A topology fingerprint approach to the inverse protein folding problem. *J. Mol. Biol.*, 227:227–238.

[Guss and Freeman, 1983] Guss, J. M. and Freeman, H. C. (1983). Structure of oxidized poplar plastocyanin at 1.6 Å resolution. *J. Mol. Biol.*, 169:521–563.

[Guss et al., 1988] Guss, J. M., Meritt, E. A., Phizackerly, R. P., Hedman, B., Murata, M., Hodgson, K. O., and Freeman, H. C. (1988). Phase determination by multiple-wavelength x-ray diffraction. Crystal structure of a basic "blue" copper protein from cucumbers. *Science*, 241:806–811.

[Hazes and Hol, 1992] Hazes, B. and Hol, W. G. J. (1992). Comparison of the hemocyanin $\beta$-barrel with other greek key $\beta$-barrels: Possible importance of the "$\beta$-zipper" in protein structure and folding. *Proteins Struct. Funct. Genet.*, 12:278–298.

[Henikoff and Henikoff, 1991] Henikoff, S. and Henikoff, J. G. (1991). Automated assembly of protein blocks for database searching. *Nucleic Acids Res.*, 19:6565–6572.

[Holm and Sander, 1993] Holm, L. and Sander, C. (1993). Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.*, 233:123–138.

[Holm et al., 1994] Holm, L., Sander, C., and Murzin, A. (1994). Three sisters, different names. *Nature Struct. Biol.*, 1:146–147.

[Jones et al., 1992] Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992). A new approach to protein fold recognition. *Nature*, 358:86–89.

[Karlin et al., 1991] Karlin, S., Bucher, P., and Brendel, V. (1991). Statistical methods and insights for protein and DNA sequences. *Annu. Rev. Biophys. Biophys. Chem.*, 20:175–203.

[Kustu et al., 1991] Kustu, S., North, A. K., and Weiss, D. S. (1991). Prokaryotic transcriptional enhancers and enhancer-binding proteins. *Trends Biochem.*, 16:397–402.

[Mathews et al., 1991] Mathews, F. S., w. Chen, Z., Bellamy, H. D., and McIntire, W. S. (1991). Three-dimensional structure of p-cresol methylhydroxylase from *pseudomonas putida* at 3.0 Å resolution. *Biochemistry*, 30:238–247.

[McIntire et al., 1986] McIntire, W., Singer, T. P., Smith, A. J., and Mathews, F. S. (1986). Amino acid sequence analysis of the cytochrome and flavoprotein subunits of p-cresol methylhydroxylase. *Biochemistry*, 20:5975–5985.

[Messerschmidt et al., 1990] Messerschmidt, A., Ladenstein, R., Huber, R., Bolognesi, M., Auigiano, L., Petrazelli, R., Rossi, A., and Finazzi-Agro, A. (1990). Refined crystal structure of ascorbate oxidase at 1.9 Å resolution. *J. Mol. Biol.*, 224:179–205.

[Muller-Hill, 1983] Muller-Hill, B. (1983). Sequence homology between lac and gal repressors and three sugar-binding periplasmic proteins. *Nature*, 302:163–164.

[Ouzounis et al., 1993] Ouzounis, C., Sander, C., Scharf, M., and Schneider, R. (1993). Prediction of protein structure by evaluation of structure-sequence fitness: Aligning sequences to contact profiles derived from three-dimensional structures. *J. Mol. Biol.*, 232:805–825.

[Pearson and Lipman, 1988] Pearson, W. R. and Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. U.S.A.*, 85:2444–2448.

[Pickett et al., 1992] Pickett, S. D., Saqi, M. A. S., and Sternberg, M. J. E. (1992). Evaluation of the sequence template method for protein structure prediction. *J. Mol. Biol.*, 228:170–187.

[Rossmann et al., 1974] Rossmann, M. G., Moras, D., and Olsen, K. W. (1974). Chemical and biological evolution of a nucleotide-binding protein. *Nature*, 250:194–199.

[Schumacher et al., 1994] Schumacher, M. A., Choi, K. Y., Zalkin, H., and Brennan, R. G. (1994). Crystal structure of LacI member, PurR, boudn to DNA: minor groove binding by $\alpha$ helices. *Science*, 266:763–770.

[Sippl and Weitckus, 1992] Sippl, M. J. and Weitckus, S. (1992). Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a database of known protein conformations. *Proteins Struct. Func. Genet.*, 13:258–271.

[Smith and Smith, 1990] Smith, R. F. and Smith, T. F. (1990). Automatic generation of primary sequence patterns from sets of related protein sequences. *Proc. Natl. Acad. Sci. U.S.A.*, 87:118–122.

[Smith and Waterman, 1981] Smith, T. F. and Waterman, M. S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.*, 147:195–197.

[Spurlino et al., 1991] Spurlino, J. C., Lu, G. Y., and Quiocho, F. A. (1991). The 2.3 Å resolution structure of the maltose- or maltodextrin-binding protein, a primary receptor of bacterial active transport and chemotaxis. *J. Biol. Chem.*, 268:5202–5219.

[Taylor, 1986] Taylor, W. R. (1986). Identification of protein sequence homology by consensus template alignment. *J. Mol. Biol.*, 188:233–258.

[Wilmanns and Eisenberg, 1993] Wilmanns, M. and Eisenberg, D. (1993). Three-dimensional profiles from residue-pair preferences: Identification of sequences with $\beta/\alpha$-barrel fold. *Proc. Natl. Acad. Sci. U.S.A.*, 90:1379–1383.

[Wodak and Rooman, 1993] Wodak, S. J. and Rooman, M. J. (1993). Generating and testing protein folds. *Curr. Opin. Struct. Biol.*, 3:247–259.

[Yi and Lander, 1993] Yi, T. M. and Lander, E. S. (1993). Protein secondary-structure prediction using nearest-neighbor methods. *J. Mol. Biol.*, 232:117–129.

[Yi and Lander, 1994] Yi, T. M. and Lander, E. S. (1994). Recognition of related proteins by iterative template refinement (ITR). *Protein Sci.*, 3:1315–1328.

## 4.8 Figures and Tables

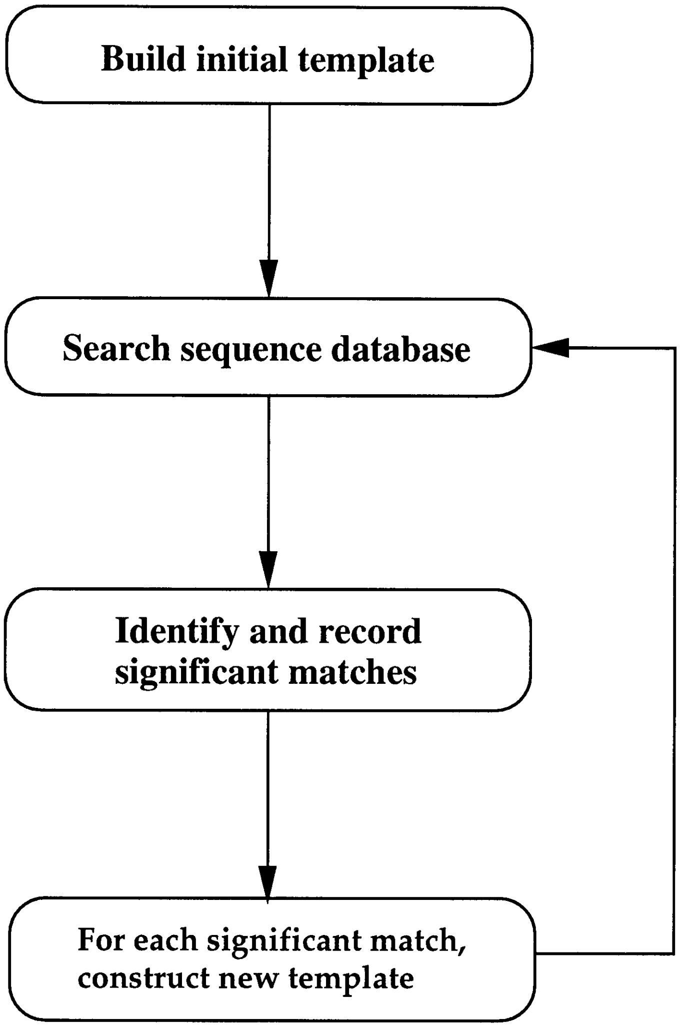

Build initial template
Search sequence database
Identify and record significant matches
For each significant match, construct new template

Figure 1A

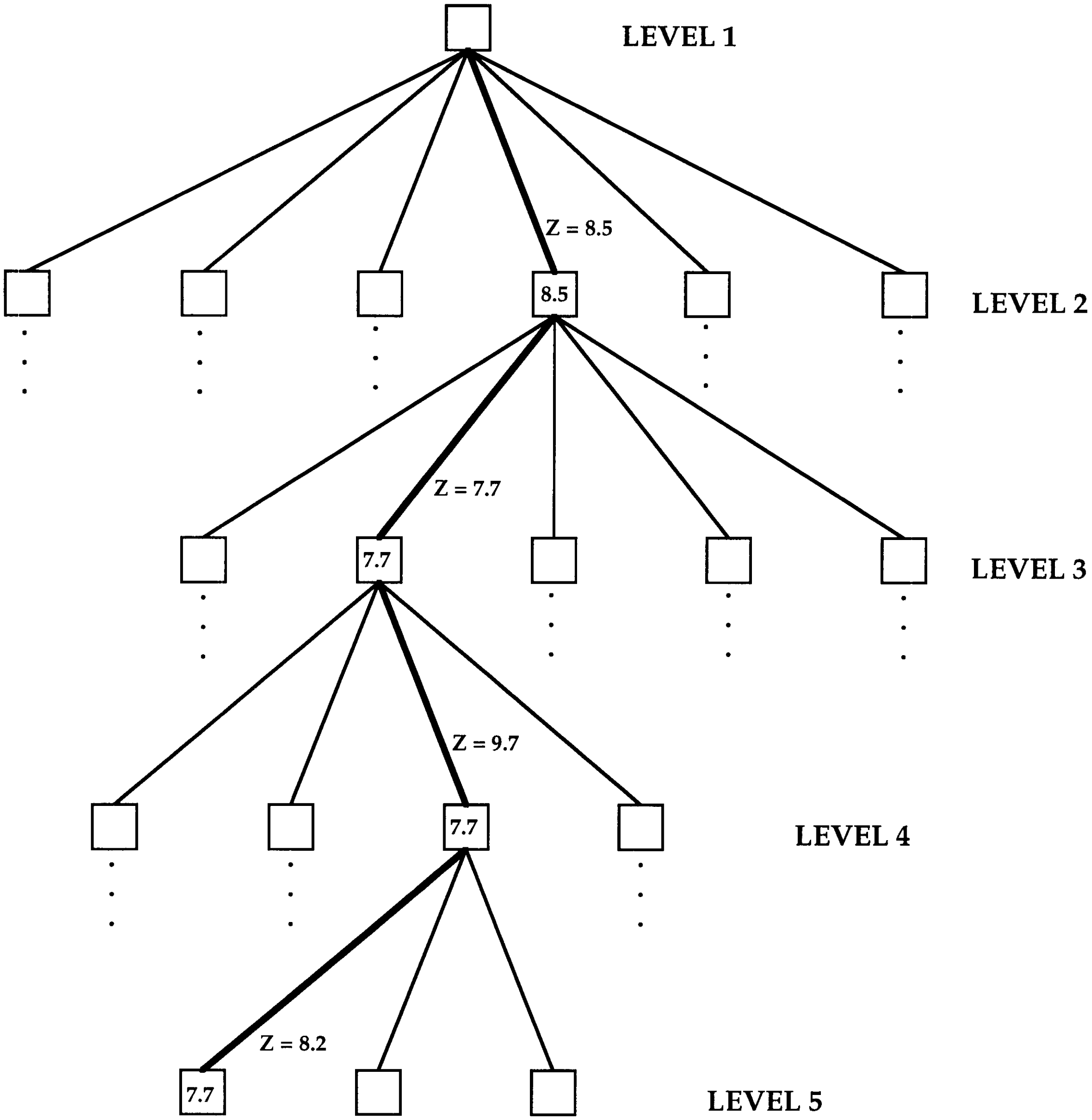

LEVEL 1
Z = 8.5
8.5
LEVEL 2
Z = 7.7
7.7
LEVEL 3
Z = 9.7
7.7
LEVEL 4
Z = 8.2
7.7
LEVEL 5

Figure 1B

## Figure 1C

```
1                                                                             80
DAA555BCDE BDB----140 0232134024 4CCAEBC978 ACE2-44024 1034024DEA DAAAAB-ACC DECAE20142     3D structure-environment profile (protein 1)

KLGFLVKQPE EPW----FQT EWKFADKAGK DLGFEVIKIA VPDG-EKTLN AIDSLAASGA KGFVIC-TPD PKLGSAIVAK     sequence (protein 1)
TIALVVSTLN NPF----FVS LKDGAQKEAD KLGYNLVVLD SQNNPAKELA NVQDLTVRGT KILLIN-PTD SDAVGNAVKM     sequence (protein 2)
TILVIVPDIC DPF----FSE IIRGIEVTAA NHGYLVLIGD CAHQNQQEKT FIDLIITKQI DGMLL---LG SRLPFDA-SI     sequence (protein 3)
AIGLVYPEND VPFNSGVFMD MVSCISRELA YHDIDLLLIA DDEH-ADCHS YMRLVESRRI DALIIAHTLD DDPRITH---     sequence (protein 4)
```

**Figure 1.** Schematic representation of Iterative Template Refinement. (A) Flow chart illustrates the iterative process of database search and template construction. (B) An expanding tree of templates grows from a starting protein during the course of the search. Each square (node) represents a distinct template. One particular path through the tree is shown in bold to illustrate the procedure for assigning Z-score labels to edges and nodes. (C) An example of a Level 4 template rooted at arabinose binding protein (1ABP) showing both the structure-environment component (represented using hexadecimal notation as in Yi and Lander, 1993) and the sequence component.

# LEVEL 1

## Template = *1ABP*

```
1
EECDAA555BCDEBDB1400 2321340244CCAEBC978A CE2440241034024DEADA AAABACCDECAE20142144 3ECCAABBBBCABEEECEDD

ENLKLGFLVKQPEEPWFQTE WKFADKAGKDLGFEVIKIAV PDGEKTLNAIDSLAASGAKG FVICTPDPKLGSAIVAKARG YDMKVIAVDDQFVNAKGKPM

101
EEAABBBABB4401441142 12400443EAEDEDBA556B BBEECDBB331131114203 44DACEECC99CEBEDCDE2 113431240044BEEBCCA5

DTVPLVMMAATKIGERQGQE LYKEMQKRGWDVKESAVMAI TANELDTARRRTTGSMDALK AAGFPEKQIYQVPTKSNDIP GAFDAANSMLVQHPEVKHWL

201
56ABBBDBB11013103EED BEDCDBAAA66BBCBB2400 EEEEECABCA65AABAADA1 320022124224EEDEAEEE CCDEECDCBBDECAEECAEE

IVGMNDSTVLGGVRATEGQG FKAADIIGIGINGVDAVSEL SKAQATGFYGSLLPSPDVHG YKSSEMLYNWVAKDVEPPKF TEVTDVVLITRDNFKEELEK

301
CEADEE
```

CEADEE 3D environment profile (*1ABP*)

KGLGGK sequence (*1ABP*)

## High-scoring proteins from search

| | Z-score | CODE | TITLE |
|---|---|---|---|
| 1 | >15.00 | JGECR | D-Ribose-binding protein precursor - Escherichia coli |
| 2 | 13.99 | JGECG | D-Galactose-binding protein - Escherichia coli |

**Figure 2A**

## LEVEL 2

### Template = *1ABP_JGECR*

```
4
DAA555BCDEBDB1400232 1340244CCAEBC978ACE2 -440241034024DEADAAA ABACCDECAE201421443E CCAABBBBCABEEECEDDEE

KLGFLVKQPEEPWFQTEWKF ADKAGKDLGFEVIKIAVPDG -EKTLNAIDSLAASGAKGFV ICTPDPKLGSAIVAKARGYD MKVIAVDDQFVNAKGKPMDT
TIALVVSTLNNPFFVSLKDG AQKEADKLGYNLVVLDSQNN PAKELANVQDLTVRGTKILL INPTDSDAVGNAVKMANQAN IPVITLDRQ--ATKG----E

103
AABBBABB440144114212 400443EAEDEDBA556BBB EECDBB33113111420344 DACEECC99CEBEDCDE211 3431240044BEEBCCA556

VPLVMMAATKIGERQGQELY KEMQKRGWDVKESAVMAITA NELDTARRRTTGSMDALKAA GFPEKQIYQVPTKSNDIPGA FDAANSMLVQHPEVKHWLIV
VVSHIASDNVLGGKIAGDYI AKKAGEGAKVIE--LQGIAG TS--AARERGEGFQQAVAAH KF--NVLASQPADFDRIKG- LNVMQNLLTAHPDVQA--VF

203                                                                                     292
ABBBDBB11013103EEDBE DCDBAAA66BBCBB2400EE EEECA-BCA65AABAADA13 20022124224EEDEAEEEC CDEECDCBBDE  3D environment
                                                                                                  profile (1ABP)
GMNDSTVLGGVRATEGQGFK AADIIGIGINGVDAVSELSK AQATG-FYGSLLPSPDVHGY KSSEMLYNWVAKDVEPPKFT EVTDVVLITRD  sequence (1ABP)
AQNDEMALGALRALQTAG-- KSDVMVVGFDG---TPDGEK AVNDGKLAATIAQLPDQIGA KGVETADKVLKGEKVQAKYP --VDLKLVVKQ  sequence (JGECR)
```

### High-scoring proteins from search

| | Z-score | CODE | TITLE |
|---|---|---|---|
| 1 | >15.00 | JGECG | D-Galactose-binding protein - - Escherichia coli |
| 2 | >15.00 | RPECDU | pur repressor - Escherichis coli |
| 3 | >15.00 | RPECL | lac represor - Escherichia coli |
| 4 | 13.95 | RPECCT | cyt repressor - Escherichia coli |
| 5 | 12.56 | RPECG | gal repressor - Escherichia coli |
| 6 | 11.51 | RPECEG | ebg repressor - Escherichia coli |
| 7 | 10.41 | JV0031 | MalI protein - Escherichia coli |
| 8 | 9.10 | A35160 | Repressor protein RafR -Escherichia coli |
| 9 | 9.04 | B24925 | lac repressor - Klebsiella pneumoniae |

Figure 2B

# LEVEL 4

## Template = *1ABP_JGECR_RPECCT_A35160*

```
4
DAA555BCDEBDB----140 02321340244CCAEBC978 ACE2-440241034024DEA DAAAAB-ACCDECAE20142 1443ECCAABBBBCABEEEC

KLGFLVKQPEEPW----FQT EWKFADKAGKDLGFEVIKIA VPDG-EKTLNAIDSLAASGA KGFVIC-TPDPKLGSAIVAK ARGYDMKVIAVDDQFVNAKG
TIALVVSTLNNPF----FVS LKDGAQKEADKLGYNLVVLD SQNNPAKELANVQDLTVRGT KILLIN-PTDSDAVGNAVKM ANQANIPVITLDRQ--ATKG
TILVIVPDICDPF----FSE IIRGIEVTAANHGYLVLIGD CAHQNQQEKTFIDLIITKQI DGMLL---LGSRLPFDA-SI EEQRNLPPMVMANEF----A
AIGLVYPENDVPFNSGVFMD MVSCISRELAYHDIDLLLIA DDEH-ADCHSYMRLVESRRI DALIIAHTLDDDPRITH--- LHKAGIPFLALGRV---PQG

98
ED-DEEAABB-BABB44014 4114212400443EAEDEDB A556BBBEECD--BB33113 111420344DACEECC99CE BEDCDE2113431240044B

KP-MDTVPLV-MMAATKIGE RQGQELYKEMQKRGWDVKES AVMAITANELD--TARRRTT GSMDALKAAGFPEKQIYQVP TKSNDIPGAFDAANSMLVQH
-----EVVSH-IASDNVLGG KIAGDYIAKKAGEGAKVIE- -LQGIAGTS----AARERGE GFQQAVAAHKF--NVLASQP ADFDRIKG-LNVMQNLLTAH
-PEL-ELPT--VHIDNLTAA FDAVNYLYE---QGHKRIG- ---CIAGPE-EMPLCHYRLQ GYVQALRRCGIMVDPQYIAR GDFTFEAG-SKAMQQLL-DL
------LPCAWFDFDNHAGT WQATQKLIAL---GHKSIA- -L--LSENT-SHSYVIARRQ GWLDALHEHGL-KDPLLRLV SP-TRRAG-YLAVMELM-SL

194                                                                                  273
EEBCCA556ABBBDBB1101 3103EEDBE---DCDBAAA6 6BBCBB2400EEEEECA-BC A65AABAADA1320022124 224E  3D environment profile
                                                                                            (1ABP)
PEVKHWLIVGMNDSTVLGGV RATEGQGFK---AADIIGIG INGVDAVSELSKAQATG-FY GSLLPSPDVHGYKSSEMLYN WVAK  sequence (1ABP)
PDVQA--VFAQNDEMALGAL RALQTAG-----KSDVMVVG FDG---TPDGEKAVNDGKLA ATIAQLPDQIGAKGVETADK VLKG  sequence (JGECR)
PQPPT-AVFCHSDVMALGAL SQAKRQGLKV--PEDLSIIG FDN----IDLTQFCDPP--L TTIAQPRYEIGREAMLLLLD QMQG  sequence (RPECCT)
PAPPT-AIITDNDLSGDGAA MALQLRG-RLSGKEAVSLVV YDGL-P-QDSIIELDVA--- AVIQSTRSLVGRQISDMVYQ IING  sequence (A35160)
```

## High-scoring proteins from search

| | Z-score | CODE | TITLE |
|---|---|---|---|
| 1 | >15.00 | RPECDU | pur repressor - Escherichis coli |
| 2 | >15.00 | RPECG | gal repressor - Escherichia coli |
| 3 | >15.00 | RPECL | lac represor - Escherichia coli |
| 4 | >15.00 | JGECG | D-Galactose-binding protein - Escherichia coli |
| 5 | >15.00 | B24925 | lac repressor - Klebsiella pneumoniae |
| 6 | >15.00 | JV0031 | MalI protein - Escherichia coli |
| 7 | >15.00 | RPECEG | ebg repressor - Escherichia coli |
| 8 | 8.03 | KIRBF | 6-Phosphofructokinase (EC 2.7.1.11) - Rabbit |
| 9 | 7.88 | S03321 | Regulatory protein nifR1 - Rhodobacter capsulatus |

**Figure 2C**

**Figure 2.** ITR with arabinose-binding protein (1ABP). The figure displays three snapshots showing the template and list of high-scoring proteins at three different stages of the search: (A) the Level 1 template, (B) one of the two Level 2 templates, and (C) a typical Level 4 template. The structure-environment string is represented using hexadecimal notation as in Yi and Lander, 1993. Below each template are the results of the database search using the template. Only significant matches are shown ($Z \geq 7.5$), and sequences closely related to one of the template sequences or to another member of the list have been removed.

# Figure 3

|  | A | C | D | E | F | G | H | I | K | L | M | N | P | Q | R | S | T | V | W | Y |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| E0 | 20 | -5 | -107 | -82 | 43 | -45 | -34 | 36 | -193 | 47 | 43 | -82 | -65 | -51 | -144 | -52 | -19 | 19 | 38 | 11 |
| E1 | 33 | 16 | -30 | -27 | -13 | 4 | -10 | 7 | -54 | 17 | 10 | -10 | -50 | -7 | 0 | -1 | -5 | 11 | -17 | -18 |
| E2 | -7 | -39 | -26 | 14 | 24 | -76 | 30 | 15 | -12 | 16 | 20 | -23 | -42 | 16 | 13 | -44 | -22 | 3 | 38 | 40 |
| E3 | 15 | -4 | -3 | 15 | -44 | -28 | -8 | -25 | 28 | -7 | -7 | -7 | -29 | 26 | 40 | -15 | 1 | -21 | -37 | -30 |
| E4 | 5 | -90 | 27 | 51 | -72 | -64 | 0 | -49 | 35 | -50 | -26 | 8 | -19 | 31 | 17 | -7 | -19 | -53 | -49 | -36 |
| E5 | -1 | 12 | -105 | -94 | 39 | -19 | -28 | 48 | -107 | 34 | 25 | -77 | -56 | -55 | -100 | -68 | -43 | 44 | 26 | 14 |
| E6 | -13 | 48 | -27 | -33 | -8 | 1 | 13 | 25 | -75 | -10 | 1 | -18 | -62 | -20 | -5 | 19 | 16 | 27 | 2 | 12 |
| E7 | -24 | -14 | -31 | -8 | 24 | -20 | 6 | 20 | -8 | 2 | -19 | -26 | -22 | -20 | 13 | -9 | 12 | 13 | 31 | 41 |
| E8 | -25 | -43 | -15 | -1 | -54 | -16 | 17 | -10 | 30 | -42 | 1 | 6 | -41 | 0 | 24 | 1 | 39 | 16 | -56 | -13 |
| E9 | -56 | -57 | -8 | 26 | -16 | -64 | -4 | -43 | 31 | -52 | -5 | 10 | -15 | 30 | 29 | 28 | 26 | -10 | -87 | -9 |
| E10 | -10 | 8 | -63 | -73 | 50 | 2 | -9 | 25 | -155 | 29 | 22 | -24 | 23 | -70 | -136 | -42 | -21 | 21 | 40 | 12 |
| E11 | -1 | 46 | 5 | -39 | -32 | 28 | 15 | -16 | -63 | -3 | -16 | 15 | 1 | -27 | -17 | 17 | 3 | -2 | -13 | -25 |
| E12 | -7 | 10 | 1 | -21 | 7 | 3 | 15 | -6 | -29 | -2 | -3 | 3 | 26 | -8 | -15 | -7 | -1 | -5 | 26 | 31 |
| E13 | 2 | 7 | 11 | -23 | -37 | 20 | 3 | -34 | 12 | -25 | -27 | 9 | 9 | -3 | 14 | 10 | 15 | -23 | -55 | -20 |
| E14 | -10 | -59 | 24 | 14 | -58 | 15 | -10 | -57 | 21 | -51 | -29 | 18 | 23 | 10 | 2 | 16 | 1 | -40 | -63 | -32 |

**Figure 3.** Structure-environment scoring matrix B. E0 to E14 denote the 15 environment classes as defined in Yi and Lander (1993). The columns are labeled with the one-letter abbreviations for the 20 amino acids.

# Table 1

## List of proteins identified by arabinose-binding protein (1ABP) search

| | PROTEIN | Z-score | Structural Similarity | Function | Ligand(s) | Family | Detected by BLAST or FASTA | Detected by Bowie-Eisenberg | Minus 3D environment profile |
|---|---|---|---|---|---|---|---|---|---|
| 1 A) | D-galactose-binding protein | 15.0 | Yes | Transport | D-galactose | periplasmic | No (0.12) | Yes (5.3) | Yes |
| B) | D-ribose-binding protein | 15.0 | Yes | Transport | D-ribose | periplasmic | No | No | Yes |
| 2 A) | PurR repressor | 15.0 | Likely | Tx. reg. | purine | lac repressor | No | No | Yes |
| B) | CytR repressor | 15.0 | Likely | Tx. reg. | cytosine | lac repressor | No | No | Yes |
| C) | GalR repressor | 15.0 | Likely | Tx. reg. | galactose | lac repressor | No | No | Yes |
| D) | LacR repressor | 15.0 | Likely | Tx. reg. | lactose | lac repressor | No | No | Yes |
| E) | MalI protein | 15.0 | Likely | Tx. reg. | maltose | lac repressor | No | No | Yes |
| F) | EbgR repressor | 15.0 | Likely | Tx. reg. | ? | lac repressor | No | No | Yes |
| G) | RafR repressor | 15.0 | Likely | Tx. reg. | raffinose | lac repressor | No | No | Yes |
| H) | RbtR operon repressor | 8.8 | Likely | Tx. reg. | ribitol | lac repressor | No | No | Yes |
| 3 | Hypothetical 125K protein | 9.7 | ? | ? | ? | ? | No | No | No |
| 4 A) | NifR1 regulatory protein | 9.6 | ? | Tx. reg. | ATP | two-component | No | No | No |
| B) | NtrC regulatory protein | 9.6 | ? | Tx. reg. | ATP | two-component | No | No | No |
| C) | DctD regulatory protein | 9.5 | ? | Tx. reg. | ATP | two-component | No | No | No |
| D) | FlbD regulatory protein | 8.4 | ? | Tx. reg. | ATP | two-component | No | No | No |
| 5 | Atrial natriuretic receptor | 9.2 | ? | Receptor | atrial nat. peptide | ? | No | No | No |
| 6 A) | Leucine-binding protein | 9.2 | Yes | Transport | leucine | periplasmic | No | Yes (5.2) | No |
| B) | LIV-binding protein | 9.2 | Yes | Transport | Leu-Ile-Val | periplasmic | No | Yes (5.6) | No |
| 7 | Succinate--CoA ligase | 9.0 | ? | Ligase | succinate, ATP, CoA | ? | No | No | No |
| 8 | Phosphofructokinase | 8.0 | Yes | Kinase | fructose-1-P, ATP | ? | No | No | No |

**Table 1.** Target proteins closely related to the starting protein or to another member of the list were removed. Remaining proteins sharing significant sequence similarity were grouped together under the same number. Structural similarity was measured by direct comparison of the crystal structures if available. Alternatively, careful sequence analysis in certain cases provided strong evidence for structural similarity with the starting protein. A brief description of the function and binding properties of each protein was provided from experimental data in the literature. Family classification was based on the family designation in the PIR sequence database or according to the recommendations reported in the references. Also indicated is whether the protein could be detected using FASTA, BLAST ($p \leq 0.05$), or the Bowie-Eisenberg method ($Z \geq 5.0$). BLAST Poisson scores less than 0.2 and Bowie-Eisenberg Z-scores greater than 4.0 are listed in parenthesis. Finally, the last column refers to the results of the search using sequence information only (i.e., the local environment profile was omitted). Question marks signify the absence of the appropriate information. 'Tx. reg.' is an abbreviation for transcriptional regulator.

# Table 2

## List of proteins identified by plastocyanin (1PCY) search

| | PROTEIN | Z-score | Structural Similarity | Function | Family | Detected by BLAST or FASTA | Detected by Bowie-Eisenberg | Minus 3D environment profile |
|---|---|---|---|---|---|---|---|---|
| 1 A) | Pseudoazurin | 15.0 | Yes | electron transfer | small copper | Yes (<0.001) | Yes (6.6) | Yes |
| B) | Amicyanin | 15.0 | Very likely | electron transfer | small copper | Yes (<0.001) | No | Yes |
| 2 | Azurin | 9.7 | Yes | electron transfer | small copper | No | No | Yes |
| 3 A) | Basic blue protein | 8.2 | Yes | electron transfer | small copper | No | No | No |
| B) | Stellacyanin | 8.2 | Very likely | electron transfer | small copper | No | No | No |
| 4 A) | Immunoglobulin light chain | 8.2 | Yes | receptor | immunoglobulin | No | Yes (6.1) | No |
| B) | Immunoglobulin heavy chain | 8.2 | Yes | receptor | immunoglobulin | No | No (4.6) | No |
| C) | T-cell receptor (α, β, δ, γ) | 8.2 | likely | receptor | immunoglobulin | No | No | No |
| D) | Class II histocompatibility antigen | 8.1 | likely | receptor | immunoglobulin | No | No | No |
| E) | T-cell surface glycoprotein (CD4, CD7, CD8) | 7.7 | Yes | receptor | immunoglobulin | No | No | No |
| 5 | Complement C2 | 8.0 | ? | protease | complement | No | No | No |
| 6 A) | L-ascorbate oxidase | 7.9 | Yes | oxidoreductase | laccase | No | No | No |
| B) | Nitrous-oxide reductase | 7.7 | ? | oxidoreductase | ? | No | No | No |

See notes for Table 1 for details.

# Table 3

## List of proteins identified by cytochrome c (1CCR) search

| | PROTEIN | Z-score | Structural Similarity | Function | Family | Detected by BLAST or FASTA | Detected by Bowie-Eisenberg |
|---|---|---|---|---|---|---|---|
| 1 | Cytochrome c551 | 9.4 | Yes | electron transfer | cytochrome c (ID) | No | No |
| 2 | Cytochrome c6 | 8.3 | likely | electron transfer | cytochrome c (IC) | No | No |
| 3 | Cytochrome c554 | 8.3 | likely | electron transfer | cytochrome c (IC) | No | No |
| 4 | Cytochrome c1aa33 | 8.1 | ? | electron transfer | cytochrome c | No | No |
| 5 | Cytochrome c555 | 8.1 | Yes | electron transfer | cytochrome c (IE) | No | No |
| 6 | p-cresol methylhydroxylase | 8.0 | Yes | oxidoreductase | cytochrome c | No | No |
| 7 | DNA polymerase (T4) | 7.7 | ? | DNA synthesis | DNA polymerase | No | Yes (6.0) |
| 8 | Cytochrome c-L precursor | 7.6 | ? | electron transfer | cytochrome c | No | No |

See notes for Table 1 for details.

# Table 4

## List of proteins identified by chymotrypsin (2CGA) search

| | PROTEIN | Z-score | Structural Similarity | Function | Family | Detected by BLAST or FASTA | Detected by Bowie-Eisenberg |
|---|---|---|---|---|---|---|---|
| 1 A) | Protease A, *S. griseus* | 9.7 | Yes | protease | Strep. serine protease | No | No |
| B) | Protease B, *S. griseus* | 9.7 | Yes | protease | Strep. serine protease | No | No |
| C) | Alpha-lytic protease | 9.6 | Yes | protease | Strep. serine protease | No | No |
| 2 | Metalloproteinase, *B. subtilis* | 9.6 | ? | protease | ? | No | No |
| 3 | Achromobacter proteinase I | 9.6 | Yes | protease | ? | No | No |
| 4 A) | Epidermolytic toxin A/B | 9.6 | ? | protease | Staph. serine protease | No | No |
| B) | Staphylococcal serine protease | 9.6 | ? | protease | Staph. serine protease | No | No |
| 5 | Hypothetical protein D-255, *M. fortuitum* | 8.2 | ? | ? | ? | No | No |
| 6 | Genome polypeptide (Rhinovirus) | 7.9 | ? | ? | Polio virus | No | No |
| 7 | Heat-shock protein HtrA | 7.5 | ? | protease | ? | No | No |

**See notes for Table 1 for details. 'Strep. serine protease' refers to the streptomyces family of bacterial serine proteases, and 'Staph. serine protease' refers to the staphylococcus family of bacterial serine proteases.**

# Table 5

## List of proteins identified by lactate dehydrogenase (5LDH, domain 1) search

| | PROTEIN | Z-score | Structural Similarity | Function | Ligand(s)/ Cofactor(s) | Family | Detected by BLAST or FASTA | Detected by Bowie-Eisenberg |
|---|---|---|---|---|---|---|---|---|
| 1 | Aspartate-semialdehyde dehydrogenase | 11.4 | ? | oxidoreductase | NAD | ? | No | No |
| 2 | GDPmannose 6-dehydrogenase | 11.0 | ? | oxidoreductase | NAD | ? | No | No |
| 3 A) | Ribitol dehydrogenase | 9.7 | ? | oxidoreductase | NAD | short-chain dehyd. | No | No |
| B) | Sorbitol-6-phosphate dehydrogenase | 9.6 | ? | oxidoreductase | NAD | short-chain dehyd. | No | No |
| C) | 27K bile acid dehydroxylating protein | 9.6 | ? | oxidoreductase | NAD | short-chain dehyd. | No | No |
| D) | Acetoacetyl-CoA reductase | 9.6 | ? | oxidoreductase | NAD | short-chain dehyd. | No | No |
| E) | Alcohol dehydrogenase (Fruit fly) | 9.6 | ? | oxidoreductase | NAD | short-chain dehyd. | No | No |
| F) | Carbonyl reductase | 9.5 | ? | oxidoreductase | NADP | short-chain dehyd. | No | No |
| 4 | UDP-N-acetylmuramoylalanine D-glutamate ligase | 9.2 | ? | ligase | ATP | ? | No | No |
| 5 A) | Transforming protein (rho-1) | 9.0 | Likely | signalling | GTP | RAS | Yes (0.02) | No (4.6) |
| B) | Transforming protein (ras) | 9.0 | Yes | signalling | GTP | RAS | No | No |
| C) | Rab4 protein | 9.0 | Likely | signalling | GTP | RAS | No | No |
| 6 | Glyceraldehyde 3-P dehydrogenase | 9.0 | Yes | oxidoreductase | NAD | ? | No | No |
| 7 | Alanine dehydrogenase | 8.8 | ? | oxidoreductase | NAD | ? | No | No |
| 8 | 3-Hydroxyacyl-CoA dehydrogenase | 8.4 | Yes | oxidoreductase | NAD | ? | No | No |
| 9 | Phosphoglycerate kinase | 8.3 | Yes | kinase | ATP | ? | No | No |
| 10 | Nitrogen fixation regulatory protein FixJ | 8.2 | ? | Tx. reg. | ATP | two-component | No | No |
| 11 | Adenylhomocysteinase | 8.1 | ? | hydrolase | ? | ? | No | No |
| 12 | Fatty-acid synthase | 8.0 | ? | oxidoreductase | NADP | ? | No | No |
| 13 | Initiation factor IF2 | 7.8 | ? | translation | GTP | EF-TU | No | No |
| 14 | Exodeoxyribonuclease V | 7.7 | ? | DNA repair | ATP | ? | No | No |
| 15 | UDPglucose 4-epimerase | 7.6 | ? | isomerase | NAD | ? | No | No |

**See notes for Table 1 for details. Only ligands and cofactors containing nucleotides are listed in "Ligand(s)/Cofactor(s)" column**

# Table 6

## List of proteins identified by tryptophan synthase (α-subunit, 1WSY_A) search

| | PROTEIN | Z-score | Structural Similarity | Function | Family | Detected by BLAST or FASTA | Detected by Bowie-Eisenberg |
|---|---|---|---|---|---|---|---|
| 1 A) | (S)-2-hydroxy-acid oxidase | 9.4 | Yes | oxidase | ? | No | No |
| B) | Flavocytochrome B2 | 9.4 | Yes | oxidase | ? | No | No |
| 2 | Indole-3-glycerol-phosphate synthase (TrpC) | 7.8 | Yes | carboxy-lyase | trp biosynthesis | No | No |
| 3 | Dihydropteroate synthase | 7.8 | ? | transferase | ? | No | No |
| 4 A) | Cyclase HisF | 7.7 | ? | cyclase | his biosynthesis | No | No |
| B) | Compound III isomerase (HisA) | 7.7 | ? | isomerase | his biosynthesis | No | Yes (5.2) |
| 5 A) | Calcium-transporting ATPase | 7.7 | ? | ion transport | ion pump ATPase | No | No |
| B) | H+-transporting ATPase | 7.7 | ? | ion transport | ion pump ATPase | No | No |
| C) | Plasma membrane ATPase I | 7.7 | ? | ion transport | ion pump ATPase | No | No |
| D) | Cadmium-transporting ATPase | 7.7 | ? | ion transport | ion pump ATPase | No | No |
| 6 | Dihydroorotate oxidase | 7.6 | ? | oxidase | ? | No | No |
| 7 | Arabinose-binding protein | 7.5 | No | transport | periplasmic | No | No |
| 8 | Transforming protein (ras) | 7.5 | No | signalling | RAS | No | No |

See notes for Table 1 for details.

# Table 7

## List of proteins of known structure identified by the searches

| Starting Protein | | Target Protein | Code | Z-score | Structural Similarity | Detected by BE, BLAST, or FASTA |
|---|---|---|---|---|---|---|
| 1. Arabinose-binding protein (1ABP) | 1a) | D-galactose-binding protein | 2GBP | 15.0 | Yes | Yes |
| | b) | D-ribose-binding protein | 2RBP | 15.0 | Yes | No |
| | 2a) | Leucine-binding protein | 2LBP | 9.2 | Yes | Yes |
| | b) | LIV-binding protein | 2LIV | 9.2 | Yes | Yes |
| | 3 | Phosphofructokinase | 1PFK | 8.0 | Yes | No |
| 2. Plastocyanin (1PCY) | 1 | Azurin | 2AZA | 9.7 | Yes | No |
| | 2 | Basic blue protein | 1CBP | 8.2 | Yes | No |
| | 3a) | Immunoglobulin light chain | 2FB4_L | 8.2 | Yes | Yes |
| | b) | Immunoglobulin heavy chain | 2FB4_H | 8.2 | Yes | No |
| | c) | T-cell surface glycoprotein, CD4 | 2CD4 | 7.7 | Yes | No |
| | 4 | L-ascorbate oxidase | 1ASO | 7.9 | Yes | No |
| 3. Cytochrome C (1CCR) | 1 | Cytochrome c551 | 351C | 9.4 | Yes | No |
| | 2 | Cytochrome c555 | --- | 8.1 | Yes | No |
| | 3 | p-cresol methylhydroxylase, cytochrome subunit | --- | 8.0 | Yes | No |
| 4. Chymotrypsin (2CGA) | 1a) | Protease A | 2SGA | 9.7 | Yes | No |
| | b) | Protease B | 2SGB | 9.7 | Yes | No |
| | c) | Alpha-lytic protease | 2ALP | 9.6 | Yes | No |
| | 2 | Achromobacter proteinase I | 1ARB | 9.6 | Yes | No |
| 5. Lactate dehydrogenase, Rossman domain (5LDH, domain 1) | 1 | Glyceraldehyde 3-P dehydrogenase | 1GD1_O | 9.0 | Yes | No |
| | 2 | p21 ras protein | 5P21 | 9.0 | Yes | No |
| | 3 | 3-Hydroxyacyl-CoA dehydrogenase | 0ACD | 8.4 | Yes | No |
| | 4 | Phosphoglycerate kinase | 3PGK | 8.3 | Yes | No |
| 6. Tryptophan synthase, α-subunit (1WSY_A) | 1a) | (S)-2-hydroxy-acid oxidase | 1GOX | 9.4 | Yes | No |
| | b) | Flavocytochrome B2 | 1FCB | 9.4 | Yes | No |
| | 2 | Indole-3-glycerol-phosphate synthase | --- | 7.8 | Yes | No |
| | 3 | Arabinose-binding protein | 1ABP | 7.5 | No | No |
| | 4 | p21 ras protein | 5P21 | 7.5 | No | No |

**Summary of data for proteins with known structures from Tables 1 - 6. Target proteins possessing significant sequence similarity were grouped together under the same number. Dashes in the 'Code' column indicate that the structure has not yet been deposited into the Brookhaven database. 'BE' refers to the Bowie-Eisenberg method.**

## Acknowledgments

We thank Bruce Tidor, Tomas Lozano-Perez, and Neil Clarke for helpful discussions and comments on the manuscript. Some of the ideas underlying ITR are related to unpublished work of E.S.L. and Robert Jones, several years ago. This work was supported in part by the National Science Foundation, the National Institutes of Health, and the Markey Foundation.

# Chapter 5

# Conclusion

This thesis has described the application of specialized database methods to three problems relating to the analysis and prediction of protein structure: the identification of structural patterns at the subdomain level, the prediction of secondary structure, and inverted protein structure prediction. In this concluding chapter, I will discuss the broader implications of these results to the field of protein folding. First, it is important to examine the relative strengths and weaknesses of the database approach and highlight some of the key technical innovations. Secondly, I shall outline promising future directions for this research. Finally, I would like to assess the relevance of this work to the Protein Folding Problem.

## 5.1 Critique of the Database Approach

The universe of possible sequences and structures is incomprehensibly large. Current data, on the other hand, suggests that sequence and structure space is only sparsely populated [Chothia, 1992, Orengo, 1994, Orengo et al., 1994]. Two possible explanations for this limited occupancy are evolutionary heritage and physico-chemical constraints. In other words, evolutionary descent from a common ancestral protein will result in a set of dense clusters of sequences and structures in sequence/structure space. Moreover, evidence is accumulating that certain topologies (and hence sequences) are favored because of the nature of the interactions that hold a protein together [Finkelstein et al., 1993, Laurents et al., 1994]. The database approach makes extensive use of these constraints in its search for meaningful patterns.

For example, the nearest-neighbor secondary structure predictor described in Chapter 3 employs a novel 'metric' for defining the neighborhood around each exemplar in the training set. A more sensitive scoring system that can detect underlying structural similarity between a test instance and a training exemplar is likely to improve prediction performance. I have used a hybrid scoring scheme that couples the Benner sequence similarity matrix to the Bowie-Eisenberg local environment method. Thus, compared to statistical and neural network techniques, the nearest-neighbor classifier makes more explicit use of evolutionary and physico-chemical relationships. It is difficult to imagine how information about amino acid substitution frequencies could be encoded in a neural network.

Likewise, I have integrated this hybrid scoring system into the ITR fold prediction technique. ITR is unusual in that it encodes both sequence and structure information into the templates, thereby combining the specificity from sequence conservation with the sensitivity of structure compatibility. The other key innovation of ITR is the iterative scheme of refining the search template with new matches from the sequence database. This process has the effect of distinguishing the signal from the noise in the template. The 'signal' positions may represent evolutionary conserved residues or amino acids that are important structural determinants. Finally, the growth of the tree of search templates (see Figure 1B in Chapter 4) reflects to a certain degree the evolutionary relationship between the seed sequence and the sequences subsequently added to the templates.

The motivation for my investigation of SSTs was the belief that there were informative structural patterns smaller than a fold but larger than supersecondary structure motifs. I have identified many such patterns, suggesting that there are potent constraints limiting the number of observed topologies at both the domain and subdomain levels. It is likely that SSTs are the product of both divergent and convergent evolution. The database approach of exhaustively comparing the entries in PDB against one another and then clustering related members into families could be extended to other levels of structural organization (i.e., supersecondary structure). In this fashion, a more systematic identification of structural patterns of all sizes could be achieved. The key technical innovation was the development of a more flexible measure of structural similarity that permitted the detection of common substructures.

What are the potential drawbacks associated with specialized database methods? I will focus on the nearest-neighbor classifier but these arguments also apply to the other database systems studied in this thesis. First, there is the issue of implementation. Devising a suitable metric and determining the composition of the database requires careful investigation. A related difficulty is that optimizing the performance of the nearest-neighbor system entails tinkering with the various parameters; there are no analytic procedures (e.g., back-propagation for neural networks) for optimizing these parameters. Moreover, it is problematic (although not impossible) incorporating more complex relationships between different features, such as the interaction between amino acids four residues apart in a helix. Finally, whereas rule-based methods can provide an explanation for why a certain peptide segment adopts a helical conformation, the output from a nearest-neighbor predictor offers no such general insight.

## 5.2 Future Directions

One natural extension of the work on SSTs is to identify in a systematic fashion recurrent structural patterns at all levels in the hierarchy of protein structural organization. Then, by creating a database of these patterns, one could develop a language for protein structure based on a vocabulary of folds, SSTs, motifs, etc. This language could be used to parse a single structure or to describe the overall universe of structures. Another important question is to what extent do SSTs correspond to autonomous folding units and folding intermediates. I have provided a few suggestive examples, but a more careful investigation may be of interest to those studying the protein folding pathway as well as protein design. Finally, one would like to better understand the physico-chemical basis for the frequent occurrence of certain structural motifs. Why are these structures favored? Close inspection of SSTs may shed some light on the underlying interactions that stabilize a particular three-dimensional conformation.

The nearest-neighbor secondary structure predictor has achieved one of the best results to date on single test sequences. The next step would be to follow the lead of Rost and Sander (1993) and use the system to predict the secondary structure of multiply aligned test sequences. Already, Salamov and Solovyev (1995) have created their own version of the nearest-neighbor classifier described in Chapter 3 and have adapted it to multiple sequence prediction. Their performance was comparable to Rost and Sander (72.2% vs. 71.6%). Other possible directions are suggested by new developments in the field of machine learning. One could divide the classifier into four expert classifiers, each of which is specialized for a specific class of proteins ($\alpha$, $\beta$, $\alpha/\beta$, $\alpha$+$\beta$). Cohen and colleagues have experimented with predicting over a single class of structures and have demonstrated improved prediction accuracy [Kneller et al., 1990]. A second more ambitious undertaking would be to attempt predicting the secondary structure of the whole sequence database, not just the proteins in PDB. A tradeoff exists with this type of unsupervised learning approach between dramatically expanding the database of exemplars versus the uncertainty surrounding the label associated with each exemplar. Finally, one can apply this methodology to the prediction of other local structural feature such as solvent accessibility which may possess more information about the tertiary fold than secondary structure.

Testing ITR on six proteins, there were 20 matches with proteins of known structure: 18 possessed the same basic fold as the seed protein, and two did not. The two false positives were instructive because both arabinose-binding protein (1ABP) and p21 Ras (5P21) possessed similar secondary structure and solvent accessibility patterns with the seed protein, the $\alpha$-subunit of tryptophan synthase (1WSY_A). Indeed, the three proteins share a common SST. One way of eliminating these spurious matches would be to add a residue-residue contact potential containing information about the relative positioning of different residues. Such a potential would complement the local environment and sequence substitution information present in the current scoring scheme. Several groups have developed contact potentials based on the loglikelihood of residue pairs being in contact [Wodak and Rooman, 1993]. More intriguingly, Gobel et al. (1994) has outlined a method for using data on the correlation of mutations in related sequences to predict residues that are in close proximity. Another area in which the sequence/structure fitness function could be improved is in the handling of deletions and insertions. Current methods employ a gap initiation and a gap extension penalty. This parameterization scheme is too restrictive to reflect the diversity in the size and number of gaps observed in related real protein structures. Finally, a more rigorous method for calculating the statistical significance of an alignment score is needed. ITR converts the raw alignment score into a Z-score, but recent theoretical results from the sequence comparison field suggests that the distribution of match scores is exponential and not normal [Karlin et al., 1991].

## 5.3 Protein Folding Problem

The protein folding problem is one of the most celebrated problems in molecular biology today. Yet, I would argue that there are two distinct subproblems embedded in this larger question. First, can one accurately predict the three-dimensional structure of a protein from its sequence? Secondly, do we understand the forces, energetics, and thermodynamics that cause a polypeptide chain to adopt a specific conformation? This thesis touches upon both issues.

Improving current methods for protein structure prediction has been a principal goal of the thesis. I have developed two promising prediction techniques: the nearest-neighbor secondary structure classifier has achieved one of the best results on single test sequences, and the ITR fold predictor was able to detect structural homology between distantly related proteins. My approach has been to exploit evolutionary and physico-chemical constraints through the use of specialized database techniques. I also made a conscious decision to work on problems associated with different levels of the protein structure hierarchy with the long-range hope of tying together the different projects. For example, the results from a secondary structure classifier could be used as input to a protein fold predictor. Similarly, local structural and contact information could be incorporated into a lattice model system to predict structure more directly. Finally, one could obtain an atomic resolution structure through homology modelling. Each of these steps requires substantial progress before *de novo* prediction can become a reality, however.

Less obvious is how this thesis has a bearing on the second question. First, I have investigated the encoding of structural information in information-based measures. A variety of local conformation, solvation, and contact potentials used in this thesis and elsewhere [Wodak and Rooman, 1993, Bowie and Eisenberg, 1993] are based on tables of information values. Exploring the relationship between these statistical potentials and physically realistic energy functions is an important research topic. Indeed, Bryant and Lawrence (1991) have estimated the dielectric constant in the interior of a protein by examining the distance distribution of charged amino acids in proteins. Secondly, prediction accuracy provides a scale for measuring the relative contribution of various factors to a particular structural phenomenon such as secondary structure formation. For example, Gibrat et al. (1991), noting that the performance of most prediction algorithms using a single sequence window as input has not exceeded 65%, have speculated that local sequence contains about 65% of the information necessary for specifying the secondary structure of a given residue. Similarly, the poor performance of methods that rely on single amino acid preferences reflects the relatively small magnitude of the intrinsic secondary structure propensities of amino acids [Minor and Kim, 1994]. Finally, my work with SSTs has furnished many examples of recurrent structural patterns. Understanding the basis for these favored packing arrangements may provide insight into the interactions that hold a protein together.

I would like to conclude the thesis with a discussion of one of the central issues in the protein structure field: Is the simple binary code of hydrophobic (buried) and hydrophilic (solvent-exposed) positions the primary determinant of the three-dimensional conformation of a protein. This question addresses both aspects of the Protein Folding Problem. Bowie et al. (1990) have matched the hydrophobicity pattern derived from a set of aligned sequences with the pattern of buried and exposed residues in protein structures in an attempt to identify the tertiary fold of the sequences. With this approach, they successfully predicted that the CheY protein is structurally similar to flavodoxin. Dill and colleagues have modelled protein folding using a lattice system consisting of only hydrophobic and polar residues [Chan and Dill, 1991]. They have found that this simple description captures many of the properties (secondary structure, compactness, unique native conformation, etc.) associated with real proteins. On the experimental side, Hecht and colleagues [Kamtekar et al., 1993] have randomized buried and exposed positions in the four-helix bundle protein cytochrome b562 with hydrophobic and hydrophilic residues, respectively. The fact that over half of these random sequences could adopt compact $\alpha$-helical structures suggests that the appropriate positioning of polar and nonpolar amino acids without regard to identity may be sufficient for inducing a particular three-dimensional topology.

The database tools developed in this thesis could assist in a more in-depth investigation of the merits and limitations of this proposal. First, an examination of related SSTs would reveal whether the underlying pattern of hydrophobic and hydrophilic positions has been conserved in domains sharing a common substructure. Furthermore, it is important to ascertain the nature of the mapping (one-to-one?) between these solvent accessibility patterns and three-dimensional topologies. Secondly, as described above, the nearest-neighbor predictor could be adapted to predict the solvent accessibility state of residues in a protein instead of their secondary structure. The nearest-neighbor approach which employs multiple window sizes would be expected to provide more information than schemes which focus on the amino acids at the predicted position only (window = 1). Thirdly, the ITR methodology could create templates combining the hydrophobicity patterns from multiple related structures or from both structures and sequences. Having information from more than one protein would facilitate the identification of the important (core) positions in the pattern. Finally, using lattice models (Yi & Lander, in progress), it may be possible to gauge the relative contribution to the native conformation of burying hydrophobic positions and exposing hydrophilic positions versus local conformational tendencies, hydrogen-bonding patterns, packing constraints, and specific (i.e., polar) contacts between different residues.

# References

[Bowie et al., 1990] Bowie, J. U., Clarke, N. D., Pabo, C. O., and Sauer, R. T. (1990). Identification of protein folds: Matching hydrophobicity patterns of sequence sets with solvent accessibility patterns of known structures. *Proteins Struct. Funct. Genet.*, 7:257–264.

[Bowie and Eisenberg, 1993] Bowie, J. U. and Eisenberg, D. (1993). Inverted protein structure prediction. *Curr. Opin. Struct. Biol.*, 3:437–444.

[Bryant and Lawrence, 1991] Bryant, S. H. and Lawrence, C. E. (1991). The frequency of ion-pair substructures in proteins is quantitatively related to electrostatic potential: A statistical model for nonbonded interactions. *Proteins Struct. Funct. Genet.*, 9:108–119.

[Chan and Dill, 1991] Chan, H. S. and Dill, K. A. (1991). Polymer principles in protein structure and stability. *Annu. Rev. Biophys. Biophys. Chem.*, 20:447–490.

[Chothia, 1992] Chothia, C. (1992). One thousand families for the molecular biologist. *Nature*, 357:543–544.

[Finkelstein et al., 1993] Finkelstein, A. V., Gutun, A. M., and Badretdinov, A. Y. (1993). Why are the same protein folds used to perform different functions? *FEBS Lett.*, 325:23–28.

[Gibrat et al., 1991] Gibrat, J. F., Robson, B., and Garnier, J. (1991). Influence of local amino acid sequence upon the zones of the torsional angles $\phi$ and $\psi$ adopted by residues in proteins. *Biochemistry*, 30:1578–1586.

[Gobel et al., 1994] Gobel, U., Sander, C., Schneider, R., and Valencia, A. (1994). Correlated mutations and residue contacts in proteins. *Proteins Struct. Funct. Genet.*, 18:309–317.

[Kamtekar et al., 1993] Kamtekar, S., Schiffer, J. M., Xiong, H., Babik, J. M., and Hecht, M. H. (1993). Protein design by binary patterning of polar and nonpolar amino acids. *Science*, 262:1680–1685.

[Karlin et al., 1991] Karlin, S., Bucher, P., and Brendel, V. (1991). Statistical methods and insights for protein and DNA sequences. *Annu. Rev. Biophys. Biophys. Chem.*, 20:175–203.

[Kneller et al., 1990] Kneller, D. G., Cohen, F. E., and Langridge, R. (1990). Improvements in protein secondary structure prediction by an enhanced neural network. *J. Mol. Biol.*, 214:171–182.

[Laurents et al., 1994] Laurents, D. V., Subbiah, S., and Levitt, M. (1994). Different protein sequences can give rise to highly similar folds through different stabilizing interactions. *Protein Sci.*, 3:1938–1944.

[Minor and Kim, 1994] Minor, D. L. and Kim, P. S. (1994). Measurement of the $\beta$-sheet-forming propensities of amino acids. *Nature*, 367:660–663.

[Orengo, 1994] Orengo, C. A. (1994). Classification of protein folds. *Curr. Opin. Struct. Biol.*, 4:429–440.

[Orengo et al., 1994] Orengo, C. A., Jones, D. T., and Thornton, J. A. (1994). Protein superfamilies and domain superfolds. *Nature*, 372:631–634.

[Rost and Sander, 1993] Rost, B. and Sander, C. (1993). Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.*, 232:584–599.

[Salamov and Solovyev, 1995] Salamov, A. A. and Solovyev, V. V. (1995). Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments. *J. Mol. Biol.*, 247:11–15.

[Wodak and Rooman, 1993] Wodak, S. J. and Rooman, M. J. (1993). Generating and testing protein folds. *Curr. Opin. Struct. Biol.*, 3:247–259.

# Acknowledgments (cont'd)

The greatest reward for enduring the hardships of graduate school has been the opportunity to meet and befriend many wonderful people. These are the people I discussed science with, played sports with, drank beer with, and had idle conversations with. I will always look fondly on my years at MIT thanks to them. In the tradition of Phil Jackson, I make the following gift of bedtime reading to my friends:

- Geno: *Coming to Terms with the Korean Male Inside of You.*
- Jim: *Becoming a Park Ranger in Five Easy Steps.*
- Jan: *Mark Morris: An Autobiography.*
- Bert: *The Complete Bill James Book on the Boston Red Sox and Rotisserie Baseball.*
- Dawn and Tony: *Big Cats and Their Prey.*
- John and Amy, Greg and Annette, Dave and Catherine, and other poker-playing buddies: *Letting Your Friends Win a Hand or Two: A Guide for Gracious Hosts.*
- Paul K.: *The Pleasures of Procrastination.*
- Andrea, Rob, and Patrick (also, Ariel and Portia): *A Jungian Analysis of the Disturbed Dreams of Siamese Cats.*
- Steve Miller: *The Glory Years of the Bulls (1991 - 1993).*
- The many people I played IM sports with at MIT: *The Joy of Losing.*
- Neil Clarke: *The Anti-Dielectic: Non-Confrontational Methods of Persuasion.*
- Ken Walsh: *1001 Experiments on Muscle Cells.*
- Todd Miller: *Darryl's Song: An Operetta in Three Acts on the Rise and Fall of Darryl Strawberry.*
- Dave Lockhart: *There is Nothing Wrong with Being Second Best.*
- Eric Schmidt: *Dream Team: The Meteoric Rise of Scottie Pippen.*
- Andy: *My Third Career as a Rock Climber* by Michael Jordan.
- Mary, Chris, and Arthur L.: *1001 Places to Throw a Margarita Party.*
- Past and present members of the Schimmel Lab (Anne, Helen, Chuong-Mong, Karin, Linda, Jonathan, Chris F., etc.): *Beyond the Second Genetic Code: The Investigation of tRNA and tRNA Synthetases.*

- Former members of the 6th floor of the Whitehead (Charo, Fay, Francisco, Wes, Mandy, etc.): *Flies, Friends, and Fun Times.*
- Peggy, Laura, and other Sive Lab members: *The Art of Sensual Tummy Massage for Frogs.*
- Dan N.: *Posing as Larry Bird's Younger Brother* by Tom Smith.
- Mark and Mary Pat: *The Best of Elevator Music from Hell.*
- Bill, Howard, Hillary and other first-generation members of the Lander Lab: *The Therapeutic Effects of Head-Butting and Lewd Jokes.*
- KK, Julois, Armando, David Wang, Jen, Johanna, Bruce, Arend, Shau Neen, Karen, Eileen and other current members of the Lander Lab: *Harmony, Discord, and Group Dynamics in the 90's.*
- Chuck: *Chuck Stories: The Zany Escapades of a Free Spirit* by Tau-Mu Yi and Jason Salter.
- Eric S.: *The Day I Woke Up: The Tales of a Reformed Pacifist.*
- Zak, Cristina, and Bob: *Automated Scripts to Fool Your Boss.*
- Past and current members of the Big Sticks (Chris Schaffner, JT, Asa, Derek, Kevin, etc.): *The Science of Hitting* by Ted Williams.
- Genome Center, Part I (Meggums, Cheryl, Chrissi, Carl, etc.): *Sequencing the Human Genome in Your Spare Time.*
- Genome Center, Part II. (Ert, Corrinne, Rob N., Lincoln, etc.): *Confessions of a Trekkie.*
- Genome Center, Part III (Angela, Mark, Bob, etc.): *Marathon for Beginners.*
- Denise, Maureen, and Eve: *The Reader's Digest Condensed Version of the NYT, Boston Globe, and Wall Street Journal.*
- Beer-drinking buddies (Scott, Keith, Jae, etc.): *Bright Lights, Fallen Angels: The Decadent Bar Scene around MIT during the Early 90's.*
- Chris and Sima: *Eating Out on a Limited Budget in Boston and the Bay Area.*
- Kei-Mu: *Reverse Psychology: Alternative Approaches to Giving Advice.*

Finally, I would like to thank Chris and Sima for their unwavering friendship, Kei-Mu for his unrelenting advice, and Mom and Dad for their love and support.