Comparative Modeling of Mainly-Beta Proteins by Profile Wrapping

by

Nathan Patrick Palmer

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Electrical Engineering / Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2006
[June 2006]

Author . . . .
Department of Electrical Engineering and Computer Science
May 24, 2006

Certified by . . . .
Bonnie A. Berger
Professor of Applied Mathematics
Thesis Supervisor

Accepted by . . . .
Arthur C. Smith
Chairman, Department Committee on Graduate Students

Comparative Modeling of Mainly-Beta Proteins by Profile Wrapping

by

Nathan Patrick Palmer

Submitted to the Department of Electrical Engineering and Computer Science
on May 24, 2006, in partial fulfillment of the
requirements for the degree of
Master of Science in Electrical Engineering / Computer Science

Abstract

The ability to predict structure from sequence is particularly important for toxins, virulence factors, allergens, cytokines, and other proteins of public heath importance. Many such functions are represented in the parallel β-helix fold class. Structure prediction for this fold is a challenging computational problem because there exists very little sequence similarity (less than 15%) across the SCOP family. This thesis introduces BetaWrapPro, a program for comparative modeling of the parallel β-helix fold. By estimating pairwise β-strand interaction probabilities, a profile of the target sequence is aligned, or "wrapped," onto an abstract supersecondary structural template. This wrapping procedure may capture folding processes that have an initiation stage followed by processive interaction between the unfolded region and the already-formed substructure. This wrap is then placed on a known structure and side-chains are modeled to produce a three-dimensional structure prediction.

We demonstrate that wrapping onto an abstract template produces accurate structure predictions for this fold (in cross-validation: average C_α RMSD of 1.55 Å in accurately wrapped regions, with 88% of the residues accurately aligned). In addition, BetaWrapPro outperforms other fold recognition methods, recognizing the β-helix fold with 100% sensitivity at 99.7% specificity in cross-validation on the PDB. One striking result has been the prediction of an unexpected parallel β-helix structure for a pollen allergen, and its recent confirmation through solution of its structure.

Thesis Supervisor: Bonnie A. Berger
Title: Professor of Applied Mathematics

Acknowledgments

The author wishes to acknowledge the National Institute of Health for its support of this work. He would also like to thank Professor Bonnie Berger, Professor Jonathan King, Professor Lenore Cowen, Phil Bradley, Andrew McDonnell and Matthew Menke for their support and guidance.

The work presented in this thesis appears in "Fold recognition and accurate sequence-structure alignment of sequences directing beta-sheet proteins," published in *Proteins: Structure, Function, and Bioinformatics*, vol. 63, pages 976–985, 2006. In addition, a poster entitled "Comparative Modeling of Mainly-β Proteins" describing this work was awarded a RECOMB and PDB Poster Award at the RECOMB 2005 conference, held in Cambridge, MA.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

Proteins are the organic polymers responsible for directing nearly all of the activity within living cells. Among various other roles, proteins serve as the basic structural components of cells and tissues, provide small molecule transport, work as antibodies to prevent infection, and act as enzymes to catalyze nearly all of the chemical reactions that take place within living cells. It is mainly through differentiation of their native *in vivo* three-dimensional shape, or *fold*, that proteins specify their particular biological function [16].

In this thesis, we present a novel computational method for the recognition and modeling of protein folds that are comprised predominantly of β-sheet secondary structure. In order to motivate our discussion of the method and illustrate results, we focus attention on the β-helix motif, a fold represented among virulence factors, allergens, toxins, autotransporters, and various other proteins of public health importance. We then show that our method is broadly applicable to various other mainly-beta protein folds.

The remainder of this chapter will present a cursory review of protein structure and the computational methods that have previously been developed for its analysis.

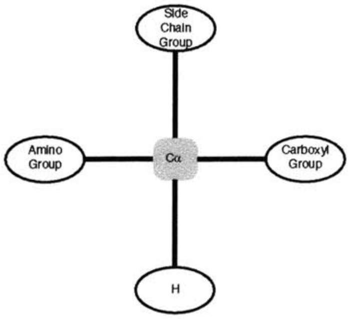

Figure 1-1: A schematic of an amino acid molecule, showing the central C_α carbon atom, the amino and carboxyl groups, the side chain group, and the hydrogen atom.

1.2 Biological Preliminaries

1.2.1 Amino Acids and Peptide Bonds

Proteins are organic macromolecules composed of multiple amino acids sequentially bound to one another to form long, flexible chains. An amino acid can, for our purposes, be viewed as having four relevant chemical components, each covalently bonded to a central carbon atom called C_α. As shown in Figure 1-1 these four chemical components are: an amino group, a hydrogen atom, a carboxyl group, and a side chain (another organic molecule that bonds to C_α at one end of its own carbon chain). There are 20 distinct amino acids, and these differ from one another only in the composition of their respective side chains – the amino, hydrogen, carboxyl, C_α composition is common to all 20.

Amino acids form covalent *peptide* (C-N) bonds with one another as the result of a dehydration synthesis reaction between the carboxyl group (COO^-) of one amino acid and the amino group (NH_3^+) of another (see Figure 1-2). After this reaction occurs, the amino donor's carboxyl group remains unaltered and exposed at one end of the new molecule (called the C-terminus), while the carboxyl donor's amino group remains at the other end (the N-terminus). The newly-bonded molecule is therefore able to accommodate an additional peptide bond at both its C and N termini (although, in nature, elongation is always observed to proceed from the N-terminus toward the

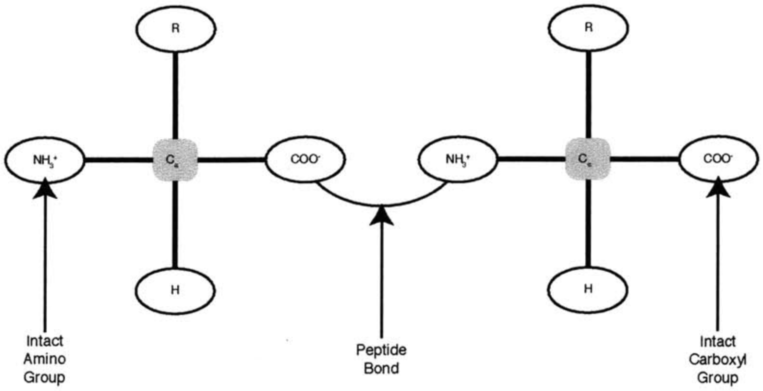

Figure 1-2: An Illustration of Peptide Bonding: Amino acids form linear chains by bonding end-to-end. The formation of the N-C peptide bond between the original amino (NH_3^+) and carboxyl (COO^-) groups results is the synthesis of a water molecule (not shown).

C-terminus), allowing the formation of extended, linear chains of amino acids.

Molecules exhibiting this repeating peptide bond structure are called *polypeptides.* Proteins are composed of one or more polypeptide chains, with each chain typically between one hundred and several thousand amino acids long. To be clear, side chain groups need not be identical for two amino acids to bond and, in fact, any of the 20 amino acids may form peptide bonds with any other. Hence, this fairly simple repeated peptide bonding scheme allows an immense amount of diversity amongst proteins, as there are 20^n different n-length polypeptides that may be formed from the naturally occurring amino acids (there are other less -frequently occurring amino acids, but they appear so infrequently that they are typically ignored by most analyses).

1.2.2 Structural Properties of Polypeptide Chains

The previous section described the basic chemistry of amino acids and polypeptide chains. We now describe the properties of a peptide bond's three-dimensional physical structure, and how these properties affect the global conformation of the polypeptide chain.

Recall that the peptide bond formed between two amino acids occurs between the CO group of one, and the NH group of the other. This covalent bond causes the resulting peptide unit to form an essentially rigid, planar structure. This planar peptide molecule is, however, able to rotate around its bonds with neighboring C_α atoms. See Figure 1-3. These peptide-C_α bonds represent the only torsional flexibility within the protein's backbone.

Thus, when studying protein structure, it is most often useful to view a polypeptide as a series of repeating peptide units (bonded CO and NH groups) connected by C_α atoms [10], rather than as a series of amino-C_α-carboxyl units bonded to one another, as in the previous section's discussion. This series of C_α-peptide-C_α-peptide-C_α-peptide... units is called the *backbone* of the protein; the ordered (from N-terminus to C-terminus) list of amino acids that are bonded together to form the polypeptide is called the protein's *sequence.*

From this perspective, each C_α atom (with the exception of the first and last in the chain) appears interposed between two peptide units, and bonded to its respective side chain group. Consequently, each amino acid is associated with two angles: one describing the rotation of the peptide plane on its amino side, called the ϕ angle, and one describing the rotation of the peptide plane on its carboxyl side, called the ψ angle (see Figure 1-4). The global conformation of a protein's backbone can therefore be completely described by specifying the ϕ and ψ angles for each of the sequence's amino acids.

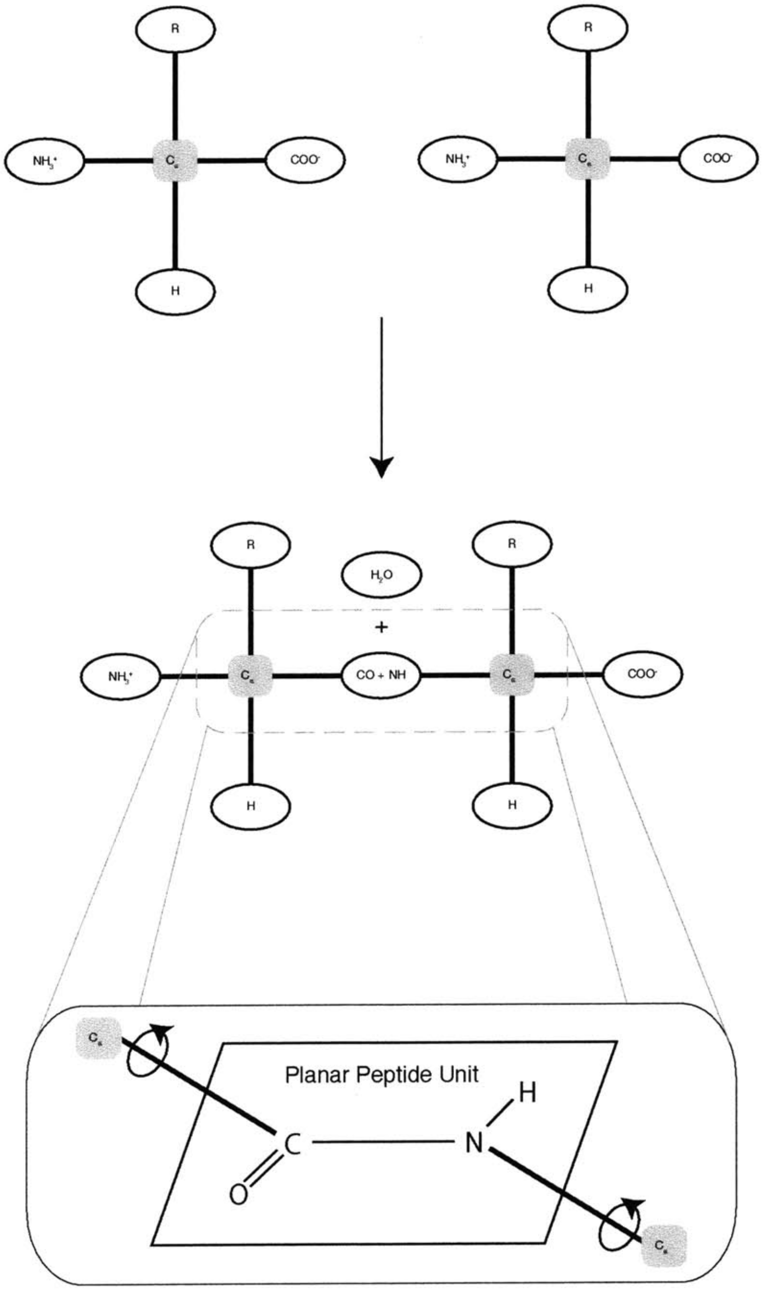

Figure 1-3: Two amino acids combine, forming a peptide bond and releasing a water molecule. The resulting peptide unit is a rigid planar structure that rotates around its bonds with neighboring C_α atoms.

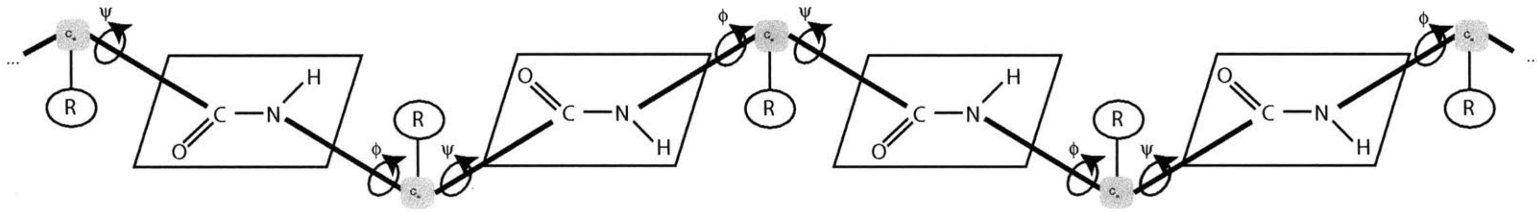

Figure 1-4: A structural view of a polypeptide chain. Each C_α atom appears between two planar peptide units – one on its amino side, and one on its carboxyl side. The peptide-C_α bonds are flexible, and the conformational angles that they assume are called the ϕ and ψ angles of a particular C_α (or, equivalently, amino acid). The side chain groups are represented in this diagram by the letter R (residue).

1.2.3 Chemical Properties of Amino Acid Side Chains

Our discussion so far has focused on the chemical properties that are *shared* amongst all of the 20 amino acids, and how those properties allow them to form elongated polypeptide chains. The last section described the basic structural properties of those elongated chains, and explained how the flexibility within each peptide-C_α bond can allow the macromolecule to assume various conformations. We now describe the *differences* between the 20 amino acids, and how these differences in chemistry govern the structure that proteins assume *in vivo.*

As noted earlier, each of the 20 amino acids, has a unique side chain group. The chemical properties of that side chain are what differentiate one amino acid from another. Table A.1 lists all 20 amino acids, grouped by their chemical properties.

There are two broad categories of chemical interactions involving the side chain groups that contribute to a protein's final structure: steric effects and weak noncovalent bonds [1].

Steric Effects

Two atoms that are not bonded to one another behave much like solid spheres, each of a fixed radius called a *van der Waals radius.* Although the the two atoms are *not* in fact solid, the repulsive force between their electron clouds increases as they move closer to each other. When they are sufficiently close in space, this repulsive force between the two atoms becomes large enough that they are effectively prevented from overlapping one another. However, a weak attractive force called a *van der Waals attraction* will cause these same two atoms to be drawn toward each other at slightly larger distances. Thus, there exists a distance of *minimum energy* between every pair of neighboring non-bonded atoms. At this distance, the magnitude of the weak attractive force is exactly equal to that of the repulsive force, while the directions of the two are exactly opposite. The consequences that result from these attractive and repulsive forces are called *steric effects* or *steric interactions.*

The fact that multiple non-bonded atoms may not occupy the same space imposes a significant constraint on the conformations that a polypeptide may assume. For example, a set of hydrophobic residues may "pack" into the internal core of a protein only if there is sufficient space in the core to accommodate a low-energy arrangement of the atoms constituting the side chains. Put another way, any such packing must be energetically favorable to leaving the residues exposed, where they will be repelled by surrounding water molecules.

Weak Noncovalent Bonds

Along with steric effects, there are several types of weak noncovalent bonds that are important to protein folding. Individually, these bonds are generally between 10 - 100 times *weaker* than the covalent bonds that couple, among other things, the CO and NH groups in a peptide bond. However, when many of these weak bonds are present, the sum of their effects can be a significant determinant of protein structure. Along with van der Waals attractions, which were described earlier, ionic bonds, hydrogen bonds and hydrophobic forces are the most important noncovalent effects that govern the spatial conformation of polypeptides.

Ionic bonds are the result of electrostatic attractions between oppositely-charged atoms. In the cell, polar water molecules tend to surround the charged ions, reducing what would otherwise be a strong attraction.

Hydrogen bonds occur when an electronegative atom attracts the electron cloud of a nearby hydrogen atom. When this attraction occurs, the hydrogen is left with a partial positive charge that then attracts a nearby electropositive atom. Thus, the electronegative and electropositive atoms become bound via the hydrogen atom shared between them. These bonds are strongest when all three atoms lie on a straight line. Since water may induce the atoms to form interfering hydrogen bonds, these bonds are considerably weaker in aqueous environments.

A significant contribution of this thesis is a study of the observed probability with which the side chains of different amino acids share hydrogen bonds with one another, both on the exposed surface and in the buried core of proteins. By learning this information from a database of known crystalized protein structures, we are able to recognize and model an important class of protein structures, given only their amino acid sequence. This problem has heretofore proven difficult to solve computationally, and our method derives a considerable amount of its power from the novel inclusion of easily-computable evolutionary information about the query sequence.

Interactions with Water

Ten amino acids (glycine, alanine, valine, leucine, isoleucine, proline, cystine, methionine, phenylalanine, tryptophan) have *nonpolar* side chains that do not interact with water. Because these residues are hydrophobic, they are typically located on the interior of proteins, where they are shielded from the aqueous environment of the cell. Cystine also plays an important roll in governing protein structure, since stabilizing disulfide bonds can form between the sulfhydryl groups of different cystine residues [16].

Five amino acids (serine, threonine, tyrosine, asparagine, glutamine) have *polar* side chains. These polar residues can form hydrogen bonds with water molecules, so are most often located on the exterior of proteins, where they interact with water in the surrounding environment [16].

Three amino acids (lysine, arginine, histidine) have charged basic side chain groups. These residues are typically positively charged in the cell (although histidine may also be neutral in vivo) and are therefore strongly hydrophilic. Like the polar residues, these three amino acids are typically found on the exterior surface of proteins where they are able to interact with water in their environment [16].

Aspartic acid and glutamic acid are the two amino acids with acidic side chains. They are therefore negatively charged within the cell, and like the basic amino acids, are driven to the exterior surface of a protein by their hydrophilic nature [16].

1.2.4 Free Energy Minimization

As Section 1.2.3 described, there are many competing forces acting on the atoms that comprise both the backbone and sidechain groups of a protein. In general, the three-dimensional conformation assumed by a protein (in the cell) is that which minimizes the free energy [1] determined by these forces. It must be the case, therefore, that all of the information required to specify a protein's structure, and hence, its function, is contained in its primary amino acid sequence. This hypothesis has been supported by a large volume of experimental evidence [1], and has given rise to the hope that a thorough understanding of the energetics involved in protein folding will someday allow computer algorithms to accurately model and predict the *in vivo* structure assumed by a protein, given only its amino acid sequence. Such a result would eliminate the need for the time-consuming and costly X-ray crystallography methods currently used to determine protein structure (see Section 1.2.5).

As described in Section 1.3.1, current methods that attempt to directly model these energy functions have met with limited success, and are not yet capable of accurately predicting the structure of molecules as large as those typically found in nature. Instead, some of the most successful state-of-the-art methods rely on guidance from statistics that capture the effect of the underlying thermodynamics without modeling them directly (see Section 1.3.2 for details). This statistics-driven approach is the general idea behind the methods presented in this thesis.

1.2.5 Experimental Determination of Protein Structure and Sequence

A protein's structure is a fundamental determinant of its biological function. Thus, in order to assemble a complete picture of how processes evolve in the cell, it is necessary to first discover the three-dimensional structure of the proteins involved. Unfortunately, this three dimensional structure is very difficult (and in some cases impossible) to determine through current experimental techniques. There are two methods that are commonly used to analyze protein structure: X-ray crystallography and nuclear magnetic resonance spectroscopy (NMR). Both are costly and time-intensive [52]. Because of this, there are relatively few "solved structures." There are several publicly-accessible databases (PDB [6], SCOP [43], ASTRAL [11]) which store and organize experimentally-derived protein structures. At the time of writing, the PDB, the canonical public database of protein structures, contained just under 35,000 solved structures.

Despite the challenges associated with structural analysis, it is relatively easy to determine a protein's primary amino acid sequence [1]. Laboratory techniques have existed since 1950 that allow researchers to do this [22], and the amount of publicly-available data has grown steadily in recent years, far outstripping the pace at which protein structures are being solved. At the time of writing, NCBI's non-redundant protein database contained roughly 3.6 billion amino acid sequences, though some of these have been inferred though the analysis of genomic data, rather than experimentation [52, 24].

The large amount of sequence data available, in light of the expense associated with structural studies, seems to justify the intense interest that has developed over the past decade in applying computational methods to the prediction of protein structure from sequence data.

1.2.6 Classifying Protein Structures

Proteins typically assume complex, irregular native folds. In order to derive a systematic method for describing protein structure through a common vocabulary, structural characteristics of proteins are usually described in terms of five levels of detail: primary sequence structure, secondary structure, supersecondary structure (folds), tertiary structure, and quaternary structure.

1.2.7 Primary Sequence Structure

A protein's primary structure is defined simply as its unbroken amino acid sequence, and is typically written down as a string of one- or three-letter amino acid codes (see Table A.1). This is, in some sense, the simplest description possible of a protein, and also the easiest feature to derive experimentally. As discussed earlier, for most proteins, it is widely believed that the composition of the amino acid sequence contains

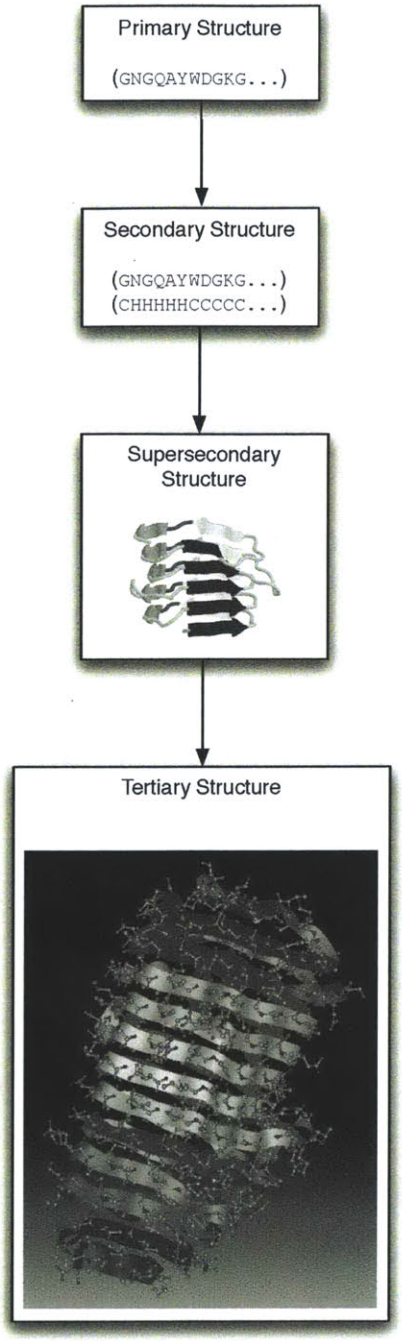

Figure 1-5: The hierarchy of protein structure detail. A protein's primary structure is simply its amino acid sequence; its secondary structure consists of a classification of each amino acid into one of a small set of commonly occuring structural subunits (e.g., `H` represents α-helical structure, `E` represents extended sheet conformation, and `C` represents coiled regions). Supersecondary structure refers to a topological arrangement of those structural subunits. Tertiary structure consists of three-dimensional coordinates for each atom in the molecule, depicted as a "ball-and-stick" diagram overlaid on the ribbon in the bottom figure.

all of the information necessary to define the three-dimensional conformation of the protein *in vivo.*

Tertiary Structure

A protein's tertiary structure is defined as a complete specification of the atomic coordinates for each atom in the molecule. Experimental methods for deriving this information are both costly and time consuming. Since proteins achieve their functionality through physical interaction with molecules in the surrounding environment, complete knowledge of a polypeptide's atomic coordinates allows a great deal of inference to be made regarding its *in vivo* function.

Secondary Structure

From a global perspective, most proteins appear to assume a highly complex shape. Viewed at a local level however, two commonly occurring patterns emerge: the α-helix and the β-sheet. The α-helix is a right-handed helix, stabilized by hydrogen bonding between the coiled rungs of the protein's backbone. The hydrogen bonding pattern induces a 3.6 residue-per-turn periodicity in the helix (see Figure 1-6). It is important to note that the stabilizing interactions within an α-helix happen at a fixed and known distance from one another. This observation gives rise to an intuitive method for searching a primary sequence for likely stretches of α-helix content in the absence of experimentally derived structure: we can simply search for short regions within the sequence where the amino acids are amenable to this hydrogen bonding pattern. For the moment we will sidestep the issue of *how* one determines whether or not a particular span of a polypeptide is likely to accommodate such a bonding pattern; we just emphasize that given the proper features to examine, we know *where* in the sequence we need look (i.e., 4 residues toward the N-terminal end and 4 toward the C-terminal end) to determine whether a particular residue is likely to be part of an α-helix.

The β-sheet is composed of multiple elongated β-strands that align with one another to form a "sheet" (see Figure 1-7). Each strand in the sheet is fully extended

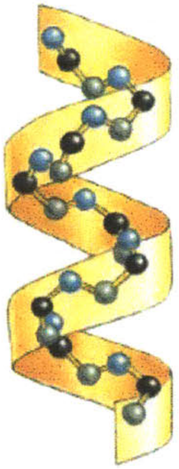

Figure 1-6: The alpha helix consists of a coiled structure where hydrogen bonds form between the consecutively stacked rungs of the helix. The bonds form between every 3.6 residues [1].

and, in this extended state, the backbones are able to hydrogen bond to one another, stabilizing the structure. β-sheets typically occur in one of two orientations: parallel or antiparallel. Parallel β-sheets consist of β-strands that are aligned with their N-to-C terminal orientation in the same direction. Antiparallel β-sheets consist of β-strands that are aligned with their N-to-C terminal orientation alternating from strand to strand (see Figure 1-7). Unlike α-helices, the β-strands that compose a β-sheet can occur a variable distance from one another in the primary amino acid sequence, making the task of computationally recognizing β-sheets significantly more challenging than in the case of α-helices, simply because we do not know "where" in the sequence to look for the next possible bonding location.

Since the strands comprising a β-sheet are fully extended, the side chain groups of the constituent amino acids lie outside the plane of the sheet. The side chains alternate orientation, with each residue pointing out of the opposite face of the sheet from its neighbors in the strand [52]. As far back as 1979, Lifson and Sander noted that certain side chain groups appear to stack preferentially against others in β-sheets [38]. Recent studies [18, 42, 52] have developed this idea in order to successfully

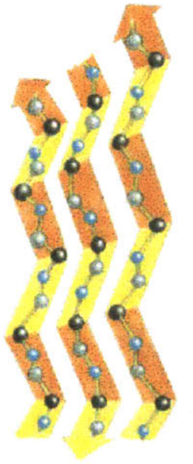

Figure 1-7: The β-sheet consists of elongated β-strands hydrogen bonded to one another through their backbones. The strands can occur a variable distance from each other in sequence, making them hard to predict computationally. This is an illustration of an anti-parallel sheet, where the N-to-C terminal orientation alternates direction between strands [1].

recognize the structure of several β-rich protein motifs from primary protein sequence data alone. This thesis will present methods derived from these techniques, and build upon their success to both improve the quality of classification and provide an estimate of full three-dimensional structures for the motifs.

While there are other less-frequently occurring structural subunits, α-helices and β-sheets are considered the most important local structure features. The less-structured regions that occur between helices and sheets are typically referred to as "random coil" or "turn" regions.

Supersecondary Structure

Frequently, tertiary structure information about a particular polypeptide is not necessary in order for biologists to make reasonable inference about the protein's function [18]. In fact, it is often the case that when two proteins share a topological arrangement of their secondary structure features, they also share common evolutionary origin and/or functionality [43]. Such a topological arrangement of secondary structure features is called a protein's supersecondary structure, and these arrangements are often referred to as protein "motifs." From a computational standpoint, this is an appealing notion, since it implies that function prediction essentially reduces to finding a solution to the supersecondary motif problem, which appears to be somewhat easier to solve.

1.2.8 Quaternary Structure

Many polypeptide chains bond to one another to form a protein complex. Often, the individual chains are inactive until forming the polymer complex. The arrangement of two or more polypeptide chains to form a complex is called quaternary structure. Quaternary structure will not be relevant to the class of proteins studied in this thesis, so we leave the discussion here, but suggest Branden and Tooze's text [10] for further reference.

1.3 Computational Preliminaries

As discussed in Section 1.2.5, the technical difficulty and expense associated with protein structure studies, combined with the large volume of sequence data, has triggered an intense interest in the development of computer algorithms that are capable of predicting important structural features of proteins given only their primary amino acid sequence. Rather than presenting an in-depth review of the technical details associated with the wide variety of computational methods available for predicting protein structure, we refer the reader to [52, 47], which present additional details. This section presents an overview of the various prediction approaches, and attempts to highlight the relative strengths of each.

1.3.1 *Ab Initio* Structure Prediction

In general, computational approaches to predicting protein structure from primary sequence data fall into one of two broad categories: those which attempt to directly simulate the folding process, and those which use statistical inference to compare a sequence to a knowledgebase of solved structures. *Ab initio* methods (literally, "from first principles") attempt to explicitly model the thermodynamics involved in the folding process [13]. The goal of these methods is to search the conformational space of a polypeptide for an arrangement of its atoms that minimizes free energy. In order to achieve a true global minimum free energy, the objective function needs to take into account every atom in the protein. Unfortunately, solving these optimization problems lies far beyond the capacity of modern computers. Several attempts have been made to leverage massively parallel super computing technology [45] (also see [55] for an overview) against molecular dynamics simulations, but rely on either simplifying assumptions about the energy function, or attempt to model only small fragments of a larger macromolecule.

One of the most successful molecular dynamics methods currently available is `Rosetta` [12]. Unlike others, it is capable of modeling large protein sequences (they report resolution of 1.5 Åor better for proteins as long.as 80 amino acids), but does so by maintaining a large database of short (3 or 4 residue) sequence fragments and their propensity to associate with one another (which have been experimentally derived). Although `Rosetta` attempts to explicitly model each local interaction and modification (at a resolution of 3-4 residues) that might be expected to occur during the folding process, its use of empirically derived scoring statistics places it somewhere between ab initio methods and comparative modeling, which we describe next.

1.3.2 Comparative Modeling

Alternatively, the *comparative modeling* approach to fold prediction sidesteps the issue of simulating the underlying physical folding process and instead takes inspiration from the field of statistical machine learning. A comparative modeling procedure compares its input sequence against a library of known sequence-structure pairs and attempts to answer the question: Is the input sequence likely to assume a similar fold to one of the proteins stored in the library?

This technique relies on a set of observations (in this case, experimentally derived protein structures) which are treated as a *training set*, or *training samples.* The goal, then, is to uncover features of the training samples that can be used to draw inference about new, previously unseen examples (e.g., an amino acid sequence whose native structure is unknown). With this information in hand, an appropriate algorithm can be devised to automatically examine its input, and respond with an answer (typically accompanied by some measure of statistical confidence) regarding the input's relationship to the training set. More broadly, the setup just described is known as a *classification problem* in the machine learning community. A complete treatment of classification paradigms is beyond the scope of this thesis, but the reader may wish to consult [51] for a thorough introduction.

An appealing feature of these statistical methods is that they need not directly account for the process that generates the observations. Thus, the issue of intractable molecular dynamics simulations disappears. Unfortunately, however, the task that we are left with is often no easier. In order to successfully design such a classification procedure, two questions need to be addressed: Which features of the data should be considered by the classifier? What is an appropriate statistical model for describing the distribution of the observed feature values in the training set? Frequently, the answers to these questions are suggested by "expert knowledge," then refined through computational methods [51].

1.3.3 Selecting an Appropriate Method

This thesis addresses the problem of identifying a class of β-rich protein motifs through this statistics-driven framework rather than through molecular dynamics simulations. In order to make an improvement over the current state-of-the-art methods for addressing this problem, we will draw upon recent insight regarding the probability with which each amino acid is likely to "stack" against any other in β-sheet structures [9, 58, 42, 52]. Note that because the strands of a β-sheet lie close in space but may be far-separated in the linear amino acid sequence of a protein, *ab initio* methods are not appropriate, since they are not generally capable of modeling long sequences.

We combine these β-sheet stacking potentials with a statistical model of evolutionary information known as sequence profiles in order to achieve greater power (and less frequent type I error, or "false positives") than the currently available methods provide.

The remainder of this section will offer a brief introduction to several of the tools currently available for protein structure analysis and prediction, since they will be referenced in later sections. Again, a more thorough treatment is available in [52] and the probabilistic underpinnings of these methods can be found in [47].

1.3.4 Sequence Homology

Recall that one of our fundamental assumptions about proteins is that their structure is determined wholly by their amino acid sequence. Thus, one of the simplest and most intuitive methods for drawing inference about a protein's structure is to search a database consisting of sequences of proteins whose structure is known, identifying those which are most similar to the sequence in question (which we will from hereon in refer to as the *query sequence*). If the query sequence is sufficiently similar (*homologous*) to any of the sequences contained in the database (we will call such a similar sequence from the database a *match* to the query), we can conclude that the query sequence also assumes a fold resembling that of the match. We need now only to define a measure of similarity between two protein sequences. We can formalize this measure by introducing the notion of *sequence alignments.*

A sequence alignment between sequences S_1 and S_2 is simply a pairing of each residue in S_1 and S_2 with either a residue from the other sequence, or a *gap*, under the constraint that if $i < j$, then s_{1_i} and s_{1_j} (the i^{th} and j^{th} characters of sequence 1) pair with s_{2_m} and s_{2_n} (the m^{th} and n^{th} characters of sequence 2) respectively such that $m < n$ or one of the pairings involves a gap. A gap character allows for explicit modeling of genomic insertion or deletion events. Each alignment has an associated score, computed as the sum of the pairwise scores for each of the paired, or *aligned* residues. The pairwise residue-residue or residue-gap alignment scores are looked up in a table containing pre-computed values which reflect the probability of observing a substitution of one amino acid for another amongst sequences which are widely accepted to be ancestrally related. The similarity measure for two sequences is therefore defined to be the score of the optimal alignment between them.

The justification for introducing gaps into an alignment derives from biologists' beliefs about the evolution of orthologous proteins. Orthologous proteins are those which share common functionality (usually between species), though are not necessarily sequence identical. It is widely believed that in the course of evolution, proteins that need to perform similar function undergo small, subtle changes, and that these changes very infrequently result in significant conformational change. Stated simply, evolutionary variation amongst orthologous proteins is thought mostly to occur at catalytic sites on a protein's surface, rather than at locations in the sequence responsible for ensuring fold stability. Thus, insertions and deletions, modeled as gaps in sequence alignments, allow for slight variation between the two sequences under the assumption that when strong similarity exists between the remaining portions of the sequences, those portions represent evolutionary conservation of structure.

There are several pairwise score matrices in common use, and we refer the reader to Durbin et al. [47] for a detailed review. The default choice for many publicly-available software packages is the symmetric BLOSUM62 matrix, derived by Henikoff and Henikoff [27]. This matrix contains log-odds values for the frequency with which each residue is paired with each other residue amongst a set of hand-aligned related sequences. The PAM matrices [20] are also frequently used and represent estimated substitution rates under the assumption that a fixed percentage of the amino acids in the sequence changed during evolutionary divergence.

Global Alignments

The problem of finding an optimal global alignment (one in which all residues must either be paired with a residue from the other sequence, or a gap) between two sequences, given such a scoring matrix, can be efficiently solved using a dynamic programming algorithm developed by Needleman and Wunsch [44]. This algorithm is simple both to understand and program, and can be easily modified to change the way that gaps are scored in order to asses a "penalty" for opening a gap if such is warranted by the context of the search.

Local Alignments

It is often times more relevant to search for short spans of highly-similar amino acid content between two sequences, since such a comparison makes no assumption of global conformational similarity. These *local alignments* are a direct extension of global alignments, but our objective is now to identify the highest-scoring alignment between any two *subsequences* of the original sequences. The result of such an alignment can shed light on whether two sequences share small fragments of common ancestral, and hence, structural similarity. Smith and Waterman [54] derived a straightforward extension of Needleman and Wunsch's algorithm intended to deal with local sequence alignments and like Needleman and Wunsch's procedure, it uses dynamic programming to find optimal alignments.

Heuristic Methods

The most common use of sequence alignment tools is to search a large database of protein sequences for matches to a query sequence. Given the large volume of sequence data currently available, the $O(n^2)$ running time of Needleman and Wunsch's procedure has become somewhat impractical for performing frequent queries. This has lead to the development of several heuristic-based search procedures which significantly improve performance at the cost of guaranteeing an optimal answer (i.e., identification of the optimal alignment). The most popular of these heuristic methods is the Basic Local Alignment Search tool, or `BLAST` [2]. In practice, `BLAST` and its successor `PSI-BLAST` achieve performance improvements that make large-scale database searching feasible while rarely failing to find optimal alignments.

Measuring Sequence Similarity and Identity

Once two sequences have been aligned, we can define two useful measures of how alike the two are:

$$Identity = \frac{Number\,of\,identical\,aligned\,residues}{Length\,of\,shorter\,sequence} \quad (1.1)$$

$$Similarity = \frac{Number\,of\,similar\,aligned\,residues}{Length\,of\,shorter\,sequence} \quad (1.2)$$

where residues having a positive alignment score in the BLOSUM62 matrix are considered similar. Proteins having greater than 25% *Identity* or greater than 40% *Similarity* values are usually assumed to have the same three-dimensional structure [48]. There are, however, many protein fold families (see [43] for an overview of hierarchical protein fold classification) that have low sequence similarity amongst their members, but share common tertiary structure. Identifying members of these families by sequence homology searches is therefore not possible. This thesis addresses the problem of identifying one such fold family, the single-stranded right-handed β-helix.

1.3.5 Profile Methods

Sequence profiles provide a method for compactly representing the amino acid content of a family of proteins. Whereas a single sequence describes the amino acid content of a single protein at each residue position, a sequence profile captures the distribution over all of the 20 possible amino acids at each position in a multiple sequence alignment. Intuitively, at the positions most responsible for maintaining fold stability, we would expect to see a good deal of sequence conservation. Consequently, the amino acid distributions at these positions in a profile end up strongly biased toward those amino acids that preserve the physical characteristics necessary to maintain stability. Gribskov et al. [26] provide a detailed overview of the motivation for employing profiles in comparative structural studies, such as this thesis. The most popular tool for creating sequence profiles is PSI-BLAST [2]. Profile-based algorithms have been successfully used to predict secondary structure in Jones' PSI-PRED software [31]. We will make use of PSI-BLAST to create sequence profiles, and PSI-PRED to help filter type I error.

1.3.6 Hidden Markov Models

Hidden Markov Models (HMMs) and their more general counterparts, Dynamic Bayesian Networks (DBNs), (see [51] for a review of probabilistic models) have become increasingly popular tools for modeling biological information. Amongst other applications, HMMs have been successfully used to perform pairwise sequence alignment [47], recognize distantly-related families of proteins [5, 21], and perform multiple sequence alignments [47]. Pfam [5] is a publicly-available database of HMMs, built from hand-curated protein sequence profiles. Each HMM corresponds to a unique family of proteins. Researchers may freely use these models to search databases for sequences that are likely to be members of the model's family. We will use an HMM-based search as a pre-processing step for the methods developed in this thesis. Durbin et al. [47] provide a thorough overview of the theory and practice of modeling biological sequence data with HMMs.

1.3.7 Threading

Threading methods operate by aligning, or "threading" a query sequence onto a library of tertiary structure templates. Frequently, these libraries simply consist of a large collection of solved protein structures. The quality of any particular threading is computed by evaluating an energy function that accounts for the local environment of each amino acid, as prescribed by the alignment onto the structural template. The optimization problem that results (identifying the minimum energy threading for a given sequence onto a particular template) is NP-complete [35], so threading schemes often estimate an initial sequence-template alignment using homology-based alignments. GenThreader [32] is one of the most commonly-used threading tools. RAPTOR [59] is a threader that was among the best-performing entrants in recent CASP competitions.

Chapter 2

BetaWrapPro

This chapter introduces a novel method, `BetaWrapPro`, for identifying protein sequences compatible with the pectin lyase-like superfamily of the single-stranded right-handed β-helix fold class, a family under the mainly-β branch of the SCOP hierarchy [43] with low sequence identity amongst its members. In addition to providing better recognition of the fold than the currently available methods, `BetaWrapPro` accurately aligns compatible sequences onto an abstract super-secondary structural template. This accurate sequence-structure alignment facilitates use of sidechain packing methods, enabling us to report estimated three-dimensional atomic coordinates for compatible sequences. We conclude with an overview of recently published results that indicate the methods presented here can be applied to other mainly-β fold classes as well.

2.1 Motivation

The *structural motif recognition* problem is: given only the target amino acid sequence for a protein, and a template for a superfamily or fold class, predict whether the protein folds into a 3-D structure which is a member of that superfamily, or fold class, or not[43]. The output of a program that performs structural motif recognition is a yes/no answer, usually accompanied by a measure of statistical confidence. The *comparative modeling* problem refers to the next step in structure prediction:

given only the target amino acid sequence for a protein, and a superfamily or fold class, predict whether the protein folds into a 3-D structure which is a member of that superfamily, or fold class; if so, give an accurate residue-by-residue alignment of the portions of the target sequence onto a super-secondary structural template, and finally, produce a prediction of the structure's atomic coordinates based on this alignment. This thesis studies the comparative modeling of a motif in a case where producing the correct sequence-target alignment was considered to be an extremely difficult problem.

Both the structural motif and the comparative modeling problems are more easily solved when there is sufficient sequence similarity between protein sequences in the superfamily, because proteins whose sequences are sufficiently similar fold into similar structures. For such a superfamily, membership queries and template alignments can be solved by simply running standard sequence matching tools such as `BLAST` and its variants[3, 2]. Even the more elaborate prediction methods, including `Threader`[32], `GenThreader`[30], and those based on hidden Markov models, rely upon structural conservation correlating to sequence conservation within the superfamily. However, there exist many protein superfamilies where, while the 3-D *structures* of the proteins are very close, there is insufficient sequence identity to determine from homology alone if an unsolved protein sequence is a member of the superfamily in question. We call such superfamilies *sequence heterogeneous.*

It has proven to be a difficult challenge to devise even structural motif recognizers for mainly-beta structures that are sequence heterogeneous. In fact, simply predicting the correct annotation of just the secondary structure of these folds can be problematic: Even the best secondary structure predictors such as `PHD`[49] and `PSIPRED`[31] more accurately predict α-helices than β-strands[36]. Insofar as general secondary structure predictors are concerned, it has been our experience that current methods do not suffice even to correctly determine the *number* of β-strands in a sequence's putative fold, much less accurately define the ends of such strands. Rather, we have found that to recognize such motifs, we must search for secondary structure and super-secondary structure *at the same time.* This was the approach taken in previous studies by Cowen *et al.*[18, 9] and Menke *et al.*[42]. Specifically, these efforts produced the `BetaWrap` program for predicting the motifs characteristic of the pectin lyase-like superfamily of the single-stranded right-handed β-helix SCOP[43] fold class, and `Wrap-and-Pack`, designed for predicting the β-trefoil motif.

The heart of both of these recognition methods is a "wrapping" algorithm[42, 18] which searches an input target sequence for aligning β-strands (parallel in the case of the β-helix, antiparallel in the case of the β-trefoil) at structurally conserved regions within the template. Whereas threading and hidden Markov model methods generally require training on representatives for each family in the superfamily (see the discussion of the performance of these methods in [18, 42]), our method is capable of accurately capturing the structural similarity across an entire superfamily with a single super-secondary structural template. Evaluation of the quality of a proposed alignment is perfomed by computing the likelihood of the alignment, given the pairwise inter-strand residue-residue correlations learned from databases of general β-strand interactions. (Both studies make a point of excluding known instances of the fold from the database used to learn their respective pairwise correlations, thus avoiding potential preferential bias for interactions that might be specific to the fold or superfamily under consideration.)

Steward and Thornton[58] take an information theory-based approach to the problem of determining the correct alignment between interacting β-strands in parallel and anti-parallel β-sheets, and their results suggest that when it is possible to limit the search to a narrow window of sequence around suspected interacting strands, consideration of inter-strand residue-residue pairings can be of significant value in determining the correct alignment between β-strands.

The purpose of this thesis is to present a program, `BetaWrapPro`, that solves the comparative modeling problem for the pectin lyase-like superfamily of the single-stranded right-handed β-helix SCOP[43] fold class (henceforth to be referred to as the β-helix motif; see Figure 2-1).

The fold is characterized by a repeating pattern of parallel β-strands in a triangular prism shape[61]. The cross-section, or *rung*, of a β-helix consists of three

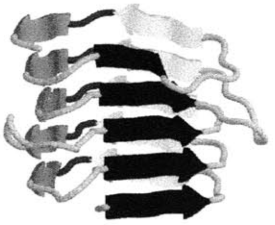

Figure 2-1: Side view of X-ray crystal structure of Pectate lyase C from Erwinia chrysanthemi[50]; β-sheet B1 is shown in light gray, B2 in medium gray, and B3 in black.

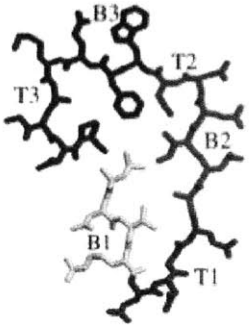

Figure 2-2: Top view of a single rung of a β-helix, parsed into β-strands B1, B2, B3 and the intervening turns T1, T2, and T3. The alternating pattern of the strands before and after T2 is conserved across the superfamily.

β-strands connected by variable-length turn regions; the backbone folds up in a helical fashion with β-strands from adjacent rungs stacking on top of each other in a parallel orientation (Figure 2-2).

2.2 Algorithm

2.2.1 Overview

`BetaWrapPro` "wraps" a profile of a target sequence onto a super-secondary structural template for a β-helix (Fig. 2-2) by searching for high-quality residue-residue interactions between aligning β-strands. A profile for a target sequence composed of n residues is simply a $20 \times n$ matrix that encodes the distribution of amino acid composition (over all 20 possible amino acids) in the columns of a multiple sequence alignment[1] performed using the target as a probe[2]. Each row of the matrix corresponds to one of the 20 amino acids, and each column corresponds to a location in the original target sequence. Thus, each entry in the matrix presents information concerning the chance of observing each amino acid aligned against a given location in the probe. Taken as a whole, this matrix presents information about residue conservation (which locations are conserved, which are variable) and what *types* of substitutions are allowed at each location[26].

Moreover, even if undetectable by straightforward homology searches across a superfamily, selective pressure on residues that stabilize a fold ought to constrain evolutionary substitution at these locations and evidence of the amino acid substitutions compatible with the fold should be present within the profile. Our method takes advantage of the fact that β-strand interactions act as a stabilizing mechanism for the β-helix by considering the potential mutations suggested by the target's profile when evaluating pairwise β-strand alignments.

We use this wrapping procedure to identify the highest scoring alignments of the target sequence onto the β-helix motif. If any of the scores derived from the wrapping procedure are sufficiently high, we then use these alignments along with a representative set of known β-helix backbone structures to perform sidechain packing. The result is a set of estimated atomic coordinates for the sequence's three-dimensional structure.

2.2.2 Details

The `BetaWrapPro` method can be divided into three distinct phases: coarse motif-compatability filtering, profile generation, and finally, alignment of the sequence/profile pair onto the motif.

. As an initial step, `BetaWrapPro` uses `BetaWrap`[18] as a subroutine to test an input target sequence for compatability with the single-stranded right-handed β-helix fold. Any sequence assigned a `BetaWrap` score below a pre-specified threshold is automatically rejected. For this study, a fairly liberal threshold of -25 (this is the

[1] Columns involving a gap in the probe are ignored.

raw score assigned by BetaWrap, which corresponds to a p-value of 0.0696, at which sensitivity is 100% and specificity is 91.1%) was chosen to ensure not only that all of the known β-helices passed the initial filtering, but also to test our method's ability to improve upon BetaWrap's specificity.

For a sequence that passes this score threshold, a profile is then created using PSI-BLAST[2] (see Sections 2.2.1, 1.3.5, and 1.3.4). A generalized version of the BetaWrap algorithm, modified to handle multiple aligned sequences and cater-corner β-sheet stacking potentials, is then used to score this profile for compatability with the β-helix fold. We describe these modifications next.

2.2.3 Generalizing the BetaWrap Algorithm

The original BetaWrap method evaluates the propensity for a polypeptide single chain to form a β-helix by computing the likelihood of inter-strand residue stacking. First, it identifies likely locations for the well-conserved B2-T2-B3 rung segment by searching for a simple hydrophobic residue sequence pattern. From each such segment, it searches forward and backward in the sequence for potential neighboring rungs that align well. The quality of any proposed rung alignment is estimated as a function of the inter-strand residue pairings that would occur from the hypothesized alignment, based on previously-derived β-sheet pairwise correlation statistics. Intuitively, the probability of a given rung alignment is the product of the probabilities (as estimated by the correlation statistics) of the individual residue-residue pairings occurring between adjacently stacked strands.

We extend this method to operate on a sequence and its accompanying profile in a natural way. Recall that the sequence's profile encodes the weight assigned to each possible residue substitution (derived from a PSI-BLAST alignment) for each position in the target sequence. Rather than compute the score for a hypothesized strand alignment to be the probability of the single chain aligning against itself, we now compute an alignment score as the product of the probability of the single-chain alignment multiplied by the weighted probability of aligning the single chain against all possible residue substitutions[2]. The weight associated with each residue substitution is simply the weight defined for that residue in the target's profile in the column corresponding to the position currently being considered in the target sequence.

The correlations used in the above computation are derived from similar β-sheets taken from the PDB, excluding the template fold class (β-helix) members. The probability that a residue of type X will align with a residue of type Y is determined by the pairwise frequency of X and Y aligning, over the frequency of X appearing, conditional on whether X is exposed or buried. Conditional probabilities are defined for stacking residues in adjacent β-strands and for cater-corner residue pairs, that is, those residues one off from a vertical alignment in either direction.

The exact formula for computing the interaction probability between positions i and j in a sequence is given in Equation 2.1, where d varies over each of the 20 amino acids, $P(r_i, d)$ is the log probability of the residue in position i interacting with d, and $f(d, j)$ is the frequency with which residue d appears in position j of the profile. The weight w assigned to the interaction is based on the relative locations of i and j: inward-pointing adjacent residues have a weight of 1, outward receive a weight of 0.5. Outward pointing adjacent residues receive a lower weight based on the assumption that structural conservation is less likely to occur on the exposed surface of a protein than in the buried core. One-off residues receive a weight of 0.25, reflecting the fact that there are twice as many one-off residues as adjacent ones.

There are several score adjustments that reflect fold-specific knowledge. A penalty of -1 is assessed to an alignment for each standard deviation from the mean number of residues between rungs. A penalty of -1 is also assessed for each large hydrophobic residue at a position that bounds one of the predicted β-strands. A bonus of +1 is granted for each pair of stacked aliphatic, aromatic, and polar residues.

$$P(i, j) = w \sum_{d=1}^{20} P(r_i, d) f(d, j) \tag{2.1}$$

[2] For efficiency and accuracy, all probabilities are actually log-transformed.

The procedure outlined above is implemented as a pruned search over all possible strand alignments, resulting in a parse of the target into high-scoring rung alignments (or wraps). Initial structural alignments of known β-helices suggested a conserved wrap size of five rungs. This rung parse induces a secondary-structure annotation (of β-strand and turn regions) on the target. Once these wraps have been generated, an α-helical secondary structure filter is applied to remove those which overlap with regions of high α-helical content (as reported by PSIPRED[31]), and the 10 top-scoring wraps are reported.

Finally, a consensus secondary structure annotation for the target is generated by breaking each of the ten best wraps into pairs of adjacent rungs and finding the four overlapping rung pairs that occur most frequently. This alignment of the target sequence to the supersecondary structural template can then be passed to standard packing methods to estimate atomic coordinates for the structurally conserved regions.

To obtain these coordinates, BetaWrapPro uses SCWRL[8] to place sidechains onto several representative backbones, and the structure with the lowest SCWRL energy score is presented in PDB format. The energy score is a measure of how well the sequence fits the backbone template: a high-energy score implies that many atoms are too close to one another, and the sequence is unlikely to form the target fold (either because it forms another fold, or it is poorly aligned to the structural template). Similar to other fold recognition programs (e.g., PROSPECT[60]), only a partial structure is output, corresponding to those portions of the template that do not include unstructured loop regions.

A publicly-available web server implementing this method is available at:
http://betawrappro.csail.mit.edu/

2.3 Methods

2.3.1 The Databases

The β-structure database was constructed from the PDB_select[29, 28] 25% list of June 2000, (with membrane proteins and the β-helices removed) as described in Cowen *et al.*, 2002[18].

The β-helix database was constructed from the sequences associated with the pectin lyase-like superfamily of the single-stranded right-handedβ-helix SCOP[43] fold class. This superfamily is comprised of eight individual families, represented by 12 unique sequences. Although there are four other superfamilies under this fold class, they each contain only one or two representative sequences. For this reason, and because several of the structures (e.g. 1hf2[17] and 1k4z, by visual inspection and PDB coordinates) do not map directly onto the generalized single-stranded right-handedβ-helix template that we consider in this study, these superfamilies were omitted from the main portion of this study.

The PDB-minus database was constructed from the amino acid sequences in RCSB's pdb_seqres database (23 June, 2004 revision), with all of those sequences represented in the β-helix database removed. This database was filtered to a 40% sequence identity non-redundant set of representatives. Low-complexity, coiled-coil, and transmembrane regions were then filtered out of this representative set. Protein sequences belonging to the leucine-rich repeat and single-stranded left-handedβ-helix SCOP fold classes were also excluded from the PDB-minus database. Members of these two classes (which correspond to the Pfam families LRR and Hexapep, resp.) conform to well-characterized sequence motifs that contain short (20-29, and 6 residue, resp.) repeats[5]. Sequences containing short repeat motifs have been experimentally observed to be a common source of false-positives generated by the `BetaWrap`[18] algorithm, and were thus excluded on the grounds of being easily filtered. Since the last revision of SCOP, four newly solved structures (PDB ids `1nhc`, `1ogm`, `1rwr`, and `1ru4`) have been identified as β-helices. These were also excluded from PDB-minus. As they have not yet been classified into a SCOP superfamily, they were also excluded from the β-helix database.

All PSI-BLAST queries were performed against the nr90 database, which was constructed from NCBI's non-redundant protein sequence database (24 June, 2004 release). This database was filtered to a 90% sequence identity non-redundant set of representatives and low-complexity, coiled-coil, and transmembrane regions were again removed.

New β-helices were identified from the SWISS-PROT[7] sequence database (Release 44.0 of 05 July, 2004), which was filtered to a 40% sequence identity non-redundant set of representatives, containing 48, 269 sequences.

Redundancy filtering was accomplished using the CD-HIT[37] program which, given an identity threshold, produces a set of representative sequences from an input database such that no two sequences in the output have greater sequence identity than the threshold value. Decreasing the amount of redundant sequence information in the PSI-BLAST search database (nr90) effectively reduced the amount of time required to build sequence profiles, while maintaining an adequate number of related sequences to construct useful alignments. The PDB-minus and SWISS-PROT databases were redundancy-filtered in order to get an accurate representation of our method's performance on the diverse sequence information present in these databases, and to avoid biasing sensitivity and specificity scores in any particular direction due to over-represented homologs.

Where appropriate, several of the databases were also filtered for low-complexity, coiled-coil, and transmembrane regions using the pfilt[31] program. Methods for identifying these regions are well-understood, and their inclusion in our databases is of little use to the present study.

2.3.2 Building Profiles

The profiles used in this study were constructed by running PSI-BLAST[2] for two iterations, and extracting the matrix of weighted observed residue frequencies reported by PSI-BLAST. All PSI-BLAST queries were performed against the nr90 database, described above, and all other parameters for PSI-BLAST were left at their default values.

2.3.3 Secondary Structure Prediction

An important part of the β-helix recognition method is screening a sequence for alpha-helical content. Whereas previously this had been achieved, primarily for simplicity's sake, by using the information-theoretic GORIV[25] method, our new algorithm uses the secondary structure predictions generated by `PSIPRED`[31]. Having already generated the `PSI-BLAST` profile for a target sequence, running `PSIPRED` presents a negligible increase in computational overhead, and allows us to leverage the information contained within the multiple sequence alignment, yielding more accurate secondary structure prediction[31]. Thus, to generate secondary structure predictions for a given sequence, we ran `PSIPRED` on the checkpoint file generated by the two iteration `PSI-BLAST` search described above, using the default number of filter iterations, with an α bias setting of 1.0 and β bias set to 1.3, as recommended by the program's authors.

2.3.4 Training

A leave-family-out cross-validation was performed on the eight β-helix families represented in the β-helix database. PDB-minus was randomly partitioned into a 60% training set and 40% testing set. For each cross, proteins in one β-helix family were placed in the testing set, while the remainder of the β-helices were placed in the training set. Parameters in `BetaWrap` are tuned according to the training set. A score threshold was then learned as the minimum score of any known β-helix contained in the training set, and this threshold was then used to classify sequences contained in the testing set.

2.4 Results

2.4.1 Recognition and Alignment of Known β-helices

On the β-helix database described in Section 2.3.1, BetaWrapPro recognizes the β-helix fold with 100% sensitivity at 99.7% specificity in cross validation. This is an improvement over the results for BetaWrap (100% sensitivity at 95.0% specificity) on the same database. This improvement in specificity resulted in over 300 sequences from the PDB-minus database that were falsely identified by BetaWrap as β-helices being correctly rejected by BetaWrapPro.

BetaWrapPro also produces accurate alignments of the target sequence onto the structural template. In sequence-heterogenous motifs such as those BetaWrapPro has been designed to predict, this is difficult to accomplish by the common homology-based sequence similarity methods. However, our profile wrapping technique proves to be successful at predicting alignment to a supersecondary structure template across diverse sequence families. All results stated in this section are from the leave-family-out cross-validation described in Section 2.3.4. In particular, we always perform sidechain packing onto a backbone taken from a family *different* from that of the target sequence.

On the 12 β-helices in our database, the sequence-structure alignment is accurate (within four position shifts of the exact position, as in Zemla et al. [62]) for 88% of predicted residues. To verify that the additional information provided by the introduction of sequence profiles assists in wrapping, BetaWrap was also modified to produce a sequence-structure alignment using the same method introduced in BetaWrapPro. The improvement was significant: BetaWrap's alignments were only 67% accurate.

2.4.2 3D Structure Prediction for the Known β-helices

The high-quality sequence-structure alignments produced by BetaWrapPro enable us to use SCWRL to generate accurate tertiary structure predictions for the conserved template motif regions. The accurately aligned regions of the β-helix template average less than 2.0 Å RMSD (see Table A.2). The sidechain predictions placed onto the backbone by SCWRL are consistent with SCRWL's reported performance on near-native backbones [8], with 61% of χ_1 and 42% of χ_{1+2} angles correct. (Dihedral angles are counted as correct if they are within 40° of the angle in the solved structure.) Table A.2 indicates that even when there is very low sequence identity between members of a fold class, the method employed by BetaWrapPro can be used to produce accurate 3D models of the proteins conforming to the fold.

2.4.3 Comparison to Other Methods

We compare BetaWrapPro to several popluar methods for fold recognition and sequence-structure alignment. Given that BetaWrapPro is specifically tailored to β-structural folds, we expect it to perform better, and this indeed turns out to be the case. As reported previously [18, 9], neither PSI-BLAST nor HMMER succeed in recognizing these folds across families. PSI-BLAST failed to recognize β-helices in a leave-family-out cross-validation (with the exceptions of the pectate lyase and pectin lyase families, and some examples across these families and the galacturonase family), and HMMER performed slightly worse.

For alignment accuracy, we compared the output of PSI-BLAST, PROSPECT Version 2 [60] and RAPTOR [59]. In all cases, we recreated the leave-family-out testing method used throughout this work, excluding from the PSI-BLAST database or the PROSPECT and RAPTOR template libraries all proteins in the same SCOP family as the target sequence. PSI-BLAST was run on the PDB database described in Section 2.3.1, filtered to 90% sequence identity. The default E-score cutoff for inclusion of 0.01 was used, and all searches converged within three rounds. PROSPECT was run using secondary structure prediction and evolutionary information from the same PDB data set just described, and z-scores were calculated with the -reliab option. RAPTOR was run with all default values.

As Table A.3 shows, BetaWrapPro produces more accurate sequence-structure alignments to the template across the entire range of the fold than more general methods. Our method finds an alignment with at least some residues correct for all but one tested sequence while the other tested programs often fail to align anything. In fact PSI-BLAST fails to produce any alignment at all on 67% of the tested structures. While PROSPECT and RAPTOR do find alignments for most of the β-helices, their alignment quality is substantially worse than BetaWrapPro's. BetaWrapPro alignments are for a 65-residue motif, while the alignments of other programs may be longer or shorter.

Finally, we mention that Liu et al. [39] have recently produced a fold recognizer tailored, like BetaWrapPro, specifically for the β-helix fold, and achieve 100% sensitivity. They report 100% specificity on an unspecified version of PDB25 and have not made their program available, so it is not possible to perform a direct comparison to BetaWrapPro.

2.4.4 Recognition of Unknown Sequences

BetaWrapPro identifies a number of putative β-helices in the SWISS-PROT data set (see Section 2.3.1). These include a number of bacterial autotransporters, including probable outer membrane proteins in Chlamydia pneumoniae (SWISS-PROT ID Q9Z813), Chlamydia muridarum (Q9PL47), and Bordetella parapertusis (P24328); AcfD from Vibrio cholerae (Q9KTQ4); adhesion and penetration protein precursor (P44596) and a putative surface exposed virulence protein from Haemophilus influenzae (P25927); and C5 epimerase from Pseudomonas aeruginosa (Q51371). We further note that while 44% of the sequences in SWISS-PROT are derived from mammals, mammalian sequences make up only 12% of the β-helix sequences that BetaWrapPro identifies from the same database, supporting earlier species distribution claims [9, 18].

BetaWrapPro also successfully identifies the newly-solved β-helical protein Jun a 1 [19] (PDB ID 1pxz), an allergen from Juniperus ashei,with a P-score of 0.0000, and filamentous hemagglutinin [14] (1rwr) from Bordetella pertussis (P-score β 0.0014), despite a lack of significant sequence identity to previously solved β-helix structures. These proteins were not included in the training set, as the structures were not available at the time.

Complete lists of high-scoring sequences detected and their `BetaWrapPro` scores can be found at the same location as our web server.

In addition to these novel predictions, Junker *et al.* [33] used `BetaWrapPro` to bootstrap a database of protein sequences that were used to identify an important class of autotransporter proteins believed to contain a β-helical passenger domain.

2.5 Discussion

Our results indicate that evolutionary information in the form of profiles generated by sequence alignments, when used in conjunction with statistics about pairwise residue-residue interactions occurring between adjacent β-strands on an abstract structural template can allow accurate fold recognition and sequence-structure alignment for the β-helix fold. Even when undetectable by straightforward sequence similarity searches across a SCOP superfamily or fold, selective pressure on residues that stabilize a fold ought to constrain evolutionary substitution at these locations and evidence of the amino acid substitutions compatible with the fold should be present within the profile. We take advantage of the fact that β-strand interactions act as a stabilizing mechanism for the β-helices by considering the potential mutations suggested by the target's profile when evaluating pairwise β-strand alignments.

We thus obtain a novel recognition and alignment method devised specifically for mainly-β supersecondary structure motifs representing sequence-heterogeneous families. The improved specificity of the `BetaWrapPro` method gives us greater confidence in the prediction of β-helices and other mainly-β folds (see Section 2.5.2). Moreover, our programs ability to produce 3D structures of newly predicted β-helices is useful in identifying novel structures, predicting functional residues, and designing mutational studies that could in turn lend support to the prediction.

We have found that sequences with a β-helix P-score of less than 0.002 have a strong likelihood of forming the fold, and those with a P-score of less than 0.01 may.

We are pursuing a number of methods to further improve the structures produced by `BetaWrapPro`, such as extending the number of rungs and attempting to pinpoint active sites. Based on X-ray scattering results [34] we also note that β-helices are a possible structure for many prions and amyloids. We hope to apply methods developed for `BetaWrapPro` to combine sequence information with CD spectral, X-ray scattering, and electron microscopy data to model and investigate the observed properties of prion tendrils.

2.5.1 Biological Implications

The majority of the β-helical protein structures deposited in the Protein Data Bank (PDB) are carbohydrate binding proteins involved in host cell recognition, infection, or penetration. One large β-helix family is the pectate lyase family, required for virulence in soft rot plant disease. Initial sequence homology studies have suggested that these proteins are representatives of a large class of virulence factors not limited to plant disease: homologous sequences are found, for example, in Yersinia pseudotuberculosis. Right-handed β-helix domains have also been found among the virulence factors and adhesins of microorganisms, including Salmonella typhimurium phage P22 [57], the plant pathogen Aspergillus niger [40], and the whooping cough pathogen Bordetella pertussis [23]. Thus the accurate prediction of β-helices in microbial or viral sequences is likely to be a useful early warning method for identifying proteins playing a role in cell attachment and penetration.

In addition, β-helices may be novel targets of anti-bacterial agents. It is known that β- helices use the lateral surface of the helix to bind polysaccharides and related molecules. We suspect that this function is particularly important for bacteria and viruses that bind to cell surfaces. Insights into the details of the mechanism of glycolysis and the specific amino acid residues involved for both lyases and hydrolases are aided by crystallographic structures of these proteins complexed with their carbohydrate substrates. The active site is located in a groove on the elongated lateral surface in the B3-T3-B1 region of the domain [56]. This general use of an elongated surface rather than a crevice is quite different from the active site clefts in conventional enzymes and may underlie the selection for this elongated fold. By locating this region in an unsolved sequence, researchers can focus their efforts on a much smaller section of the protein. Since parallel β-helices appear to be relatively rare in mammals and humans, such inhibitors may be very specific for their protein substrates. In addition, since they will be sugar analogues, solubility and transport problems should be relatively easy to overcome.

The prediction and subsequent confirmation of the β-helical conformation of a pollen allergen is of particular interest. Given the role of β-helices as microbial virulence factors and toxins, it is not surprising that the immune system mounts an efficient response to these proteins. It may be that where a plant pollen surface protein has evolved a β-helical fold for dealing with polysaccharide metabolism, the immune system responds as if to a microbial pathogen.

2.5.2 Application to Other Structures

In recently published work [41], we show that the method of wrapping profiles onto supersecondary structural templates developed in this thesis can be successfully applied to other mainly-β fold classes. In particular, we show that for the β-trefoil fold, another sequence-heterogeneous SCOP family, our method recognizes fold instances with 100% sensitivity at 92.5% specificity. Additionally, `BetaWrapPro` was able to correctly align 89% of the residues in conserved structural positions (see Table A.4). Again, as in the case of the β-helix, previously existing methods do not perform as well as `BetaWrapPro` on either the recognition or 3D modeling tasks for this fold (see Table A.5).

The β-trefoil SCOP superfamilies that `BetaWrapPro` was tested on include the interleukin-1 cytokines, promoters of mammalian immune system response, appetite regulation [46], and insulin secretion [15], among other functions. The same superfamily contains the fibroblast growth factors, important for cell growth and differentiation. The STI-like superfamily includes neurotoxins produced by both Clostridium tetani and Clostridium botulinum,and homologues of Salmonella typhimurium-derived T-cell inhibitor, which acts to subvert the hosts immune response to invaders [4]. Rather than acting as enzymes like the β-helices, the β-trefoils interact with receptor proteins to induce specific behaviors in a cell.

As in Section 2.4.4, we used `BetaWrapPro` to search the SWISS-PROT database for novel predictions of the β-trefoil fold. Among putative β-trefoils identified are the Kunitz-type proteinase inhibitor BbCI (P83051) from Bauhinia bauhinioides, agglutinin-like protein 3 (P46590) from Candida albicans, and protein B17 (P33878), a proposed virulence factor in the smallpox virus [53].

As with the β-helices, the ability to recognize and model the β-trefoils has various uses for the scientific community, including working towards a better understanding of the structure- function relationship of new trefoils. This may be useful for structural studies, identifying new trefoils, and drug discovery and design.

In addition, our results suggest that the methods presented in this thesis may be broadly applicable to predicting the structure of a wide variety of mainly-β folds from amino acid sequence alone. In order to realize this goal, it will first be necessary to automate the process of constructing the supersecondary structural templates onto which the wrapping algorithm aligns an input sequence (the templates used in this thesis were hand-curated). With such a tool in hand, the methods presented in this thesis should provide a sound framework for a wide variety of comparative modeling-based protein structural studies.

2.6 Conclusion

Our results indicate that the profile wrapping method developed in this thesis can be successfully applied to recognize and model the β-helical fold and make novel structure predictions for sequences whose fold is unknown. We have shown that evolutionary information, compactly represented by multiple sequence alignment-derived profiles, can be integrated with pairwise β-sheet residue stacking potentials in order to produce high-quality alignments of amino acid sequences onto supersecondary structural templates. As our ability to automatically generate abstract templates improves, these methods may eventually be used to predict structure for many more families of mainly-β proteins, one of the most difficult protein structure prediction problems that remains.

Appendix A

Tables

Amino Acid	**Three-Letter Code**	**One-Letter Code**	**Chemical Property**
Glycine	Gly	G	Nonpolar
Alanine	Ala	A	Nonpolar
Valine	Val	V	Nonpolar
Leucine	Leu	L	Nonpolar
Isoleucine	Ile	I	Nonpolar
Proline	Pro	P	Nonpolar
Cystine	Cys	S	Nonpolar
Methionine	Met	M	Nonpolar
Phenylalanine	Phe	F	Nonpolar
Tryptophan	Trp	W	Nonpolar
Serine	Ser	S	Polar
Threonin	Thr	T	Polar
Tyrosine	Tyr	Y	Polar
Asparagine	Asn	N	Polar
Glutamine	Gln	Q	Polar
Lysine	Lys	K	Basic
Arganine	Arg	R	Basic
Histidine	His	H	Basic
Aspartic acid	Asp	D	Acidic
Glutamic acid	Glu	E	Acidic

Table A.1: The 20 amino acids and their abbreviations, grouped by the chemical properties of their side chain groups.

Structure	P-score	Alignment	C_αRMSD	χ_1	χ_{1+2}	Aligned sequence	BLAST
(PDB code)		accuracy	(Å)	correct	correct	identity (%)	E-score
1air	0.0000	1.00	1.21	0.63	0.23	18	0.06
1bn8	0.0009	1.00	1.31	0.62	0.43	22	0.06
1ee6	0.0032	1.00	1.70	0.55	0.24	14	0.04
1jta	0.0000	1.00	1.45	0.71	0.32	12	0.02
1bhe	0.0000	0.80	0.97	0.67	0.34	19	0.03
1czf	0.0005	1.00	1.43	0.72	0.46	11	0.18
1rmg	0.0002	0.94	1.29	0.54	0.36	17	0.07
1dab	0.0000	1.00	1.39	0.60	0.41	18	1.0
1dbg	0.0000	1.00	2.83	0.66	0.45	14	0.02
1qjv	0.0012	0.66	2.13	0.45	0.29	11	0.82
1idk	0.0014	1.00	1.47	0.55	0.35	16	2.8
1h80	0.0015	0.14	1.40	0.67	0.50	11	0.23
Average:	0.0023	0.88	1.55	0.61	0.42	16	

Table A.2: Cross-Validation Score and Modeling Accuracy for the β-helix Structures, as Packed onto the Minimum-Energy Template Structure from Outside its Own SCOP Family. Families are separated by a single line. The P-Score is the `BetaWrapPro` score for the sequence. RMSD, dihedral angle correctness, and aligned sequence identity are only calculated on the accurately aligned residues of the structure. Aligned sequence identity is the identity between the query sequence and the template structure it was aligned to by `BetaWrapPro`. BLAST E-Score is the expectation score bl2seq [2] gives to its best alignment between the query and template sequences.

Structure	**BetaWrapPro**		**PSI-BLAST**		**PROSPECT**		**RAPTOR**	
	Aligned exactly	Aligned within 4	Aligned exactly	Aligned within 4	Aligned exactly	Aligned within 4	Aligned exactly	Aligned within 4
1air	0.88	1.00	0.76	0.95	0.78	0.90	0.87	0.95
1bn8	0.94	1.00	0.78	0.78	0.88	0.95	0.88	0.95
1ee6	1.00	1.00	none	none	0.29	0.32	0.35	0.43
1jta	0.94	1.00	0.74	0.75	0.74	0.80	0.87	0.81
1idk	0.94	1.00	0.77	0.81	0.74	0.83	0.88	0.94
1bhe	0.80	0.80	none	none	none	0.08	0.50	0.58
1czf	1.00	1.00	none	none	0.24	0.34	0.54	0.58
1rmg	0.94	0.94	none	none	0.38	0.59	0.50	0.58
1dab	1.00	1.00	none	none	0.35	0.37	none	none
1dbg	0.94	1.00	none	none	0.22	0.37	0.59	0.67
1h80	none	0.14	none	none	0.20	0.43	none	0.25
1qjv	0.66	0.66	none	none	0.20	0.27	none	0.17
total:	0.91	0.88	0.76	0.81	0.41	0.49	0.72	0.62

Table A.3: Percent of Sequence-Structure Alignment Correct for Several Programs on the β-helices. An entry of "none" indicates that no residues were correctly aligned.

Structure (PDB code)	P-score	Alignment accuracy	C_αRMSD (Å)	χ_1 correct	χ_{1+2} correct	Aligned sequence identity (%)	BLAST E-score
1bff	0.111	1.00	3.79	0.52	0.39	10	0.011
1g82	0.041	1.00	6.23	0.56	0.50	7	2.6
2i1b	0.012	1.00	5.59	0.47	0.22	13	1.1
1irp	0.011	1.00	4.12	0.36	0.34	12	0.077
2ila	0.010	1.00	5.65	-	-	7	no hits
1wba	0.038	1.00	3.58	0.51	0.55	7	0.99
1tie	0.041	1.00	2.51	0.46	0.35	15	1.7
3bta	0.013	0.08	-	-	-	13	3.2
1a8d	0.009	none	-	-	-	5	0.23
Average:	0.038	0.89	4.50	0.48	0.39	10	-

Table A.4: Cross-Validation Score and Modeling Accuracy for the β-trefoil Structures, as Packed onto the Minimum-Energy Template Structure from Outside its Own SCOP Family. Superfamilies are separated by a single line. The P-Score is the `BetaWrapPro` score for the sequence. RMSD, dihedral angle correctness, and aligned sequence identity are only calculated on the accurately aligned residues of the structure. Aligned sequence identity is the identity between the query sequence and the template structure it was aligned to by `BetaWrapPro`. BLAST E-Score is the expectation score bl2seq [2] gives to its best alignment between the query and template sequences. Note that the structure 2ila in the PDB does not include sidechain coordinates. We do not report structure prediction results for 3bta because only four residues are accurately aligned.

Structure	BetaWrapPro		PSI-BLAST		PROSPECT		RAPTOR	
	Aligned exactly	Aligned within 4	Aligned exactly	Aligned within 4	Aligned exactly	Aligned within 4	Aligned exactly	Aligned within 4
1bfg	0.33	1.00	0.15	0.45	0.22	0.87	0.38	1.00
1g82	0.17	1.00	none	none	0.35	0.57	none	none
21bi	0.33	1.00	0.15	0.45	0.48	0.57	none	0.40
1irp	0.33	1.00	none	none	none	none	0.75	1.00
2ila	0.33	1.00	none	none	0.47	0.73	0.40	0.62
1wba	0.50	1.00	none	none	0.08	0.42	none	1.00
1tie	058	1.00	none	none	0.23	0.68	0.25	0.58
3bta	none	0.08	none	none	none	none	none	none
1a8d	none	none	none	none	none	0.03	none	none
total: 0.30 0.89 0.15 0.45 0.30 0.77 0.45 0.78								

Table A.5: Percent of Sequence-Structure Alignment Correct for Several Programs on the β-trefoils. An entry of "none" indicates that no residues were correctly aligned.

Bibliography

[1] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell.* Garland Publishing, New York, fourth edition, 2002.

[2] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI–BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 25:3389–3402, 1997.

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

[4] T. Arai and K. Matsui. A purified protein from salmonella typhimurium inhibits high-affinity interleukin-2 receptor expression on ctll-2 cells. *FEMS Immunol. Med. Microbiol.*, 17:155–160, 1997.

[5] A. Bateman, L. Coin, R. Durbin, R.D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E.L.L. Sonnhammer, D.J.Stud holme, C. Yeats, and S.R. Eddy. The pfam protein families database. *Nucleic Acids Res.*, 32:D138–D141, 2004.

[6] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The protein data bank. *Nucleic Acids Res.*, 28:235–242, 2000.

[7] B. Boeckmann, A. Bairoch, R. Apweiler, M.C. Blatter, A. Estreicher, E. Gasteiger, M.J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, and

M. Schneider. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, 31:365–370, 2003.

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

[9] P. Bradley, L. Cowen, M. Menke, J. King, and B. Berger. BETAWRAP: Successful prediction of parallel β-helices from primary sequence reveals an association with many microbial pathogens. *Proc Natl Acad Sci, USA*, 98:14819–14824, 2001.

[10] C. Branden and J. Tooze. *Introduction to Protein Structure.* Garland Publishing, New York, 1991.

[11] S. E. Brenner, P. Koehl, and M. Levitt. The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Research*, 28:254–256, 2000.

[12] C. Bystroff and D. Baker. Prediction of local structure in proteins using a library of sequence-structure motifs. *Journal of Molecular Biology*, 281:565–577, 1998.

[13] D. Chivan, T. Robertson, R. Bonneau, and D. Baker. Ab initio methods. *Methods of Biochemical Analysis*, 44:547–557, 2003.

[14] B. Clantin, H. Hodak, E. Willery, C. Locht, F. Jacob-Dubuisson, and V. Villeret. The crystal structure of filametous hemaglutinin secretion domain and its implication for the two-partner secretion pathway. *Proc. Natl. Acad. Sci. USA*, 101:6194–6199, 2004.

[15] P.G Comens, B.A. Wolf, E.R. Unanue, P.E. Lacy, and M.L. McDaniel. Interleukin 1 is potent modulator of insulin secretion from isolated rat islets of langerhans. *Diabetes*, 36:963–970, 1987.

[16] G. M. Cooper. *The Cell, A Molecular Approach.* ASM Press, Washington, D.C., second edition, 2000.

[17] S.C. Cordell, R.E. Anderson, and Jan Löwe. Crystal structure of the bacterial cell division inhibitor minc. *The EMBO Journal*, 20:2454–2461, 2001.

[18] L. Cowen, P. Bradley, M. Menke, J. King, and B. Berger. Predicting the beta-helix fold from protein sequence data. *Journal of Computational Biology*, 9:261–276, 2002.

[19] E. W. Czerwinski, T. Midoro-Horiuti, M. A. White, E. G. Brooks, and R. M. Goldblum. Crystal structure of jun a 1, the major cedar pollen allergen from juniperus ashei, reveals a parallel beta-helical core. *J. Biol. Chem.*, 280:3740–3746, 2005.

[20] M. O. Dayhoff and R. M. Schwartz. Matrices for detecting distant relationships. *Atlas of Protein Sequence and Structure*, 5, sup. 3:353–358, 1978.

[21] S. R. Eddy. Profle hidden markov models. *Bioinformatics*, 14(9):755–763, 1998.

[22] P. Edman. Protein sequence determination. *Acta Chem. Scand.*, 4:283–299, 1950.

[23] P. Emsley, I.G. Charles, N.F. Fairweather, and N.W. Isaacs. Structure of bordetella pertussis virulence factor p.69 pertactin. *Nature*, 381:90–92, 1996.

[24] E.S. Lander et al. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

[25] J. Garnier, J. Gibrat, and B. Robson. Gor secondary structure prediction method version iv. *Methods Enzymol.*, 266:540–553, 1996.

[26] M. Gribskov, R. Lüthy, and D. Eisenberg. Profile analysis. *Methods in Enzymology*, 183:146–159, 1990.

[27] S. Henikoff and J. G. Henikoff. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA*, 89:10915–10919, 1992.

[28] U. Hobohm and C. Sander. Enlarged representative set of protein structures. *Protein Sci.*, 3:522–524, 1994.

[29] U. Hobohm, M. Scharf, R. Schneider, and C. Sander. Selection of a representative set of structures from the brookhaven protein data bank. *Protein Sci.*, 1:409–417, 1992.

[30] D.T. Jones. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.*, 287:797–815, 1999.

[31] D.T. Jones. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, 292:195–202, 1999.

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

[33] M. Junker, C. Schuster, A. McDonnell, K. Sorg, M. Finn, B. Berger, and C. Clark. Pertactin beta-helix folding mechanism suggests common themes for the secretion and folding of autotransporter proteins. *Proceedings of the National Academy of Sciences*, 103.

[34] A. Kishimoto, K. Hasegawa, H. Suzuki, H. Taguchi, K. Namba, and M. Yoshida. Beta-helix is a likely core structure of yeast prion sup35 and amyloid fibers. *Biochem. Biophys. Res. Commun.*, 315:739–745, 2004.

[35] R. Lathrop. The protein threading problem with sequence amino acid interaction preference is np-complete. *Protein Engineering*, 7:1059–1068, 1994.

[36] A.M. Lest, L. Lo Conte, and T.J.P. Hubbard. Assessment of novel fold targets in casp4: predictions of three-dimensional structures, secondary structures and interresidue contacts. *Proteins*, Suppl 5:98–118, 2001.

[37] W. Li, L. Jaroszewski, and A. Godzik. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics*, 17:282–283, 2001.

[38] S. Lifson and C. Sander. Antiparallel and parallel [beta]-strands differ in amino acid residue preferences. *Nature*, 282:109–111, 1979.

[39] Y. Liu, J. Carbonell, P. Weigele, and V. Gopalakrishna. Segmentation conditional random fields (scrfs): a new approach to protein fold recognition. In *Proceedings of the ninth annual international conference on Computational molecular biology*, pages 408–422, 2005.

[40] O. Mayans, M. Scott, I. Connerton, T. Gravesen, J. Benen, J. Visser, R. Pickersgill, and J. Jenkns. Two crystal structures of pectin lyase a from aspergillus reveal a ph driven conformational change and striking divergence in the substrate-binding clefts of pectin and pectate lyases. *Structure*, 5:677–689, 1997.

[41] A. McDonnell, M. Menke, N. Palmer, J. King, L. Cowen, and B. Berger. Fold recognition and accurate sequence-structure alignment of sequences directing beta-sheet proteins. *Proteins: Structure, Function, and Bioinformatics*, 63.

[42] M. Menke, E. Scanlon, J. King, B. Berger, and L. Cowen. Wrap-and-pack: a new paradigm for beta structural motif recognition with application to recognizing beta trefoils. In *Proceedings of the eighth annual international conference on Computational molecular biology*, pages 298–307. ACM Press, 2004.

[43] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, 247:536–540, 1995.

[44] S. Needleman and C. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48:443–453, 1970.

[45] V. S. Pande, I. Baker, J. Chapman, S. P. Elmer, S. Khaliq, S. M. Larson, Y. M. Rhee, M. R. Shirts, C. D. Snow, E. J. Sorin, and B. Zagrovic. Atomistic protein folding simulations on the submillisecond timescale using worldwide distributed computing. *Biopolymers*, 68:91–109, 2003.

[46] C.R. Plata-Salaman. Meal patterns in response to the intracerebroventricular administration of interleukin-1 beta in rats. *Physiol. Behav.*, 55:727–733, 1994.

[47] R.Durbin, S.Eddy, A. Krough, and G. Mitchinson. *Biological Sequence Analysis.* Cambridge University Press, New York, NY, 2003.

[48] B. Rost. Twilight zone of protein sequence alignments. *Protein Engineering*, 12(2):85–94, 1999.

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

[50] B. Rost and C. Sander. Conservation and prediction of the solvent accessibility in protein families. *Proteins*, 20:216–226, 1994.

[51] S. Russel and P. Norvig. *Artificial Intelligence, A Modern Approach.* Pearson Education, Inc., Upper Saddle River, NJ, second edition, 2003.

[52] E. L. Scanlon. Predicting the triple beta-spiral fold from primary sequence data. Master's thesis, Massachusetts Institute of Technology, 2004.

[53] S. N. Shchelkunov, V. M. Linov, and L. S. Sandakhchiev. Genes of variola and vaccina viruses necessary to overcome the host protective mechanism. *FEBS Lett*, 319:80–83, 1993.

[54] T. Smith and M. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147:195–197, 1981.

[55] C. D. Snow, H. Ngyen, V. S. Pande, and M. Gruebele. Absolute comparison of simulated and experimental protein-folding dynamics. *Nature*, 420:102–106, 2002.

[56] S. Steinbacher, U. Baxa, S. Miller, A. Weintraub, R. Seckler, and R. Huber. Crystal structure of phage p22 tailspike protein complexed with salmonella sp. *o-antigen receptors. Proc. Natl. Acad. Sci. USA*, 93:10584–10588, 1996.

[57] S. Steinbacher, R. Seckler, S. Miller, B. Steipe, R. Huber, and P. Reinemer. Crystal structure of p22 tailspike protein: interdigitated subunits in a thermostable trimer. *Science*, 265:383–386, 1994.

[58] R.E. Steward and J.M. Thornton. Prediction of strand pairing in antiparallel and parallel beta-sheets using information theory. *Proteins*, 48(2):178–191, 2002.

[59] J. Xu, D. Kim, and Y. Xu. RAPTOR: the optimal protein threading by linear programming. *J. Bioinform. Comp. Biol.*, 1:95–117, 2003.

[60] Y. Xu and D. Xu. Protein threading using PROSPECT: design and evaluation. *Proteins*, 40:343–354, 2000.

[61] M. Yoder, S. Lietzke, and F. Jurnack. Unusual structural features in the parallel beta-helix in pectate lyases. *Structure*, 1:241, 1993.

[62] A. Zemla, C. Venclovas, J. Moult, and K. Fidelis. Processing and analysis of CASP3 protein structure predictions. *Proteins*, 37(S3):22–29, 1999.