Predicting the Beta-Trefoil Fold from Protein Sequence Data

by

Matthew Ewald Menke

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 and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2004

Author . . . . . .
Department of Electrical Engineering and Computer Science
January 16, 2004

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

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

Predicting the Beta-Trefoil Fold from Protein Sequence Data

by

Matthew Ewald Menke

Submitted to the Department of Electrical Engineering and Computer Science
on January 16, 2004, in partial fulfillment of the
requirements for the degree of
Master of Science in Electrical Engineering and Computer Science

Abstract

A method is presented that uses β-strand interactions at both the sequence and the atomic level, to predict the beta-structural motifs in protein sequences. A program called `Wrap-and-Pack` implements this method, and is shown to recognize β-trefoils, an important class of globular β-structures, in the Protein Data Bank with 92% specificity and 92.3% sensitivity in cross-validation. It is demonstrated that `Wrap-and-Pack` learns each of the ten known SCOP β-trefoil families, when trained primarily on β-structures that are not β-trefoils, together with 3D structures of known β-trefoils from outside the family. `Wrap-and-Pack` also predicts many proteins of unknown structure to be β-trefoils. The computational method used here may generalize to other β-structures for which strand topology and profiles of residue accessibility are well conserved.

Thesis Supervisor: Bonnie Berger
Title: Professor of Applied Mathematics

Acknowledgments

I would like to thank Professor Bonnie Berger for her invaluable supervision of my work and the writing of this thesis. Special thanks goes to Professor Lenore Cowen of Tufts University for her insightful ideas and suggestions.

I would also like to acknowledge Professor Johnathon King for his help with understanding the biology behind the problem. Eben Scanlon provided help with numerous aspects of the work.

I am greatful for the funding provided by the Merck fellowship awarded to me by the MIT Computational and Systems Biology Initiative.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Hand-curated hierarchical classification systems, such as SCOP [18] and CATH [19] divide the set of known globular protein folds into groups depending on the overall topology of the fold. At the top level of the hierarchy, protein folds are divided into such classes as "mainly alpha," "mainly beta," "alpha/beta," etc. according to the predominant type of secondary structure motifs in the protein; the classification specializes within these classes into fold, superfamily, and family levels of the hierarchy. The *structural motif recognition* problem is the following prediction problem: given only the amino acid sequence for a protein, and a target fold or superfamily, predict whether the protein folds into a 3D structure which is a member of that fold, or superfamily, or not.

The structural motif recognition problem is more easily solved when there is sufficient sequence similarity between protein sequences in the target fold, because proteins whose sequences are sufficiently similar fold into similar structures. For such a fold, membership questions can be solved by simply running standard sequence matching tools such as `BLAST` [1]. However, there exist many protein folds where while the 3D *structures* of the proteins are very close, there is insufficient sequence similarity to determine from sequence alone if an unsolved protein sequence is a member of the target fold. Such folds are called such folds *sequence heterogeneous.*

It has proved to be a difficult challenge to devise structural motif recognizers for mainly-beta structures that are sequence heterogeneous. In fact, even the best (local)

secondary structure predictors are better at correctly placing α-helices than β-strands [20, 23]. It has been our experience that general secondary structure predictors do not suffice even to correctly determine the *number* of β-strands in a sequence that folds into one of these motifs; never mind find the ends of the strands accurately. 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 by our first structural motif recognizer for a sequence-heterogeneous mainly-beta fold, `BetaWrap`, which predicts the right-handed parallel β-helix motif [7, 6, 10].

`BetaWrap` used a structural template approach to look for the conserved elements of super-secondary structure in the β-helix motif. Given a sequence of unknown structure, the program would "wrap" the sequence into a parallel β-helix with conserved regions of β-strands and variable-length turn regions. For each possible wrap, pairwise statistical preferences of which amino acids prefer to stack on top of each other in the β-sheets was calculated, and compared to a database of stacking preferences in amphipathic β-sheets (from general non-β-helix β-structural motifs that had such β-sheets). This approach can detect more global interactions than a local secondary structure predictor, allowing for the capture of relationships between residues that are close in space, but may be far, and an irregular distance apart, in sequence. With some additional complexity, such as adding a secondary structure filter to rule out false positives with too much global α-helical content, `BetaWrap` was able to completely separate the true β-helices from the non-β-helices in a 2000 non-redundant version of the PDB, in a leave-family-out cross validation.

The purpose of this paper is to introduce a new method for solving structural motif recognition problems that arise in sequence-heterogeneous β-structural superfamilies, that we call `Wrap-and-Pack`. As the name indicates, the method we employ has two phases, a wrapping phase and a packing phase. The wrapping phase is conceptually similar to what `BetaWrap` does for the β-helices: it tries to parse the structure onto an abstract template that captures the conserved elements of super-secondary structure, and screens for favorable pairwise correlations between adjacent residues in the putative β-sheets; It also incorporates bonuses and penalties into the score, such as

the β-propensity of residues in the putative β-strands, according to PSIPRED. When we create a wrapping phase and apply it alone to the β-trefoil fold, we find that it does fairly well at identifying the correct regions of conserved secondary structure in the true β-trefoils; however, unlike BetaWrap and the β-helices, there are non-trefoil sequences that the program indicates could form β-trefoils. To help screen these out, we go on to the packing phase.

The packing phase incorporates, for the first time, 3D energetic information into our structural template by way of a backbone dependent rotamer library [21]. In particular, the most favorable wraps are fed into the SCWRL program of Canutescu et al., which then threads the wrap onto a small set of β-trefoil backbones, resolves steric clashes, and reports an energy score. The energy score is used to help discriminate the trefoils from the non-trefoils.

Another way to think of Wrap-and-Pack is as a two-tiered threader. A threader is a method that tries to map a sequence of unknown structure onto the backbone of a known 3D structure (see [9, 14, 22] and [17] for a recent survey). Threaders perform best when the two sequences have very similar 3D structures, and additionally, when the correct conserved portions of the unknown sequence are mapped to the conserved portions of the known structure. In this case, correct threadings should produce low energy scores. In the case where there is high sequence similarity, sequence alignment methods can generate the appropriate mapping. However, in the sequence-heterogeneous case, general threading methods typically must explore a large search space, trying many different ways to thread the sequence to the structure. In addition to the computational issues, this can lead to false positives, both because the energy scores are only approximations, and also false positive sequences can score well by chance in some possible threading, when the number of allowed different threadings is large. Both the wrapping phase and the packing phase of Wrap-and-Pack can be seen as threaders– the first with a motif-recognition style energy function based on pairwise residue correlations with no molecular dynamics. Then the small number of high-scoring alignments of the target sequence wrapped onto the template structure are passed to a threader that uses the more sensitive 3D packing molecular dynamics

constraints (in our case, we use SCWRL on a conserved portion of the trefoil cap strand, but other threaders could also be substituted). The key is by using a template and an initial structure-based wrapping phase, we can drastically reduce the number of different ways each sequence is threaded onto the target backbone in the packing phase. Then, if one of these small number of sequence threadings produces a low-energy score using SCWRL, as we did, or some other 3D sidechain placement algorithm, we are more confident that it is a true positive example.

We demonstrate the utility of the Wrap-and-Pack method on the motif recognition problem for the β-trefoil fold [18] (see Figure 1-1). The β-trefoil consists of three leaves around an axis of three-fold symmetry. Each of the leaves consists of four β-strands, B1, B2, B3, and B4, separated by turn regions T1, T2, and T3. The three B2-T2-B3 β-hairpins form a cap on one end of the barrel (bottom of figure). The cap strands B2 and B3 have slightly twisted backbones. The B1 and B4 strands of all three leaves form a six-stranded antiparallel β-barrel. Like β-helices, the β-strands have one side exposed to the environment of the cell and one side buried within the protein; however, the stacked rungs of the β-helices tend to be much more uniform than the β-strands within a trefoil leaf. Furthermore, only three residues of adjacent barrel strands are aligned because of the sheer of the barrel.

There are 122 solved β-trefoil proteins according to the current SCOP version 1.63 in six superfamilies, and ten families. The β-trefoils serve as neurotoxins, inhibitors, and receptors. β-trefoil proteins have also been implicated in inducing the inflammatory response in rheumatoid athritis, as well as playing roles in embryonic development, and tumorigenesis.

Wrap-and-Pack recognizes β-trefoils with 92% specificity and 92.3% sensitivity on a non-redundant version of the PDB in a leave-family-out cross validation (see Sections 4 and 3.2). The Wrap-and-Pack program also identifies a large set of sequences as having strong β-trefoil potential when run on the databases SWISS-PROT ([2], see Section 4). Table 4.1 lists a few of the top-scoring proteins of unsolved structure that we predict to have a β-trefoil structure based on their performance using Wrap-and-Pack.

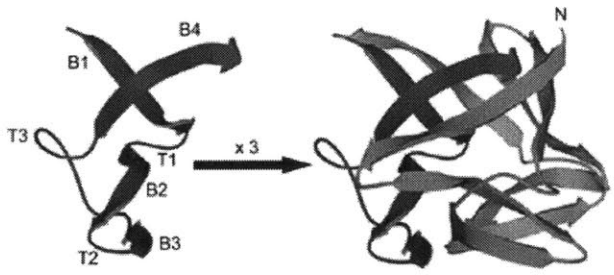

Figure 1-1: The β-trefoil consists of three leaves around an axis of three-fold symmetry. In this figure a single leaf is shown in dark gray (left). Each of the leaves consists of four β-strands, B1, B2, B3, and B4, separated by turn regions T1, T2, and T3. B2-T2-B3 form a β-hairpin. T1 and T2 both contain β-turns. The B1 and B4 strands of all three leaves form a six-stranded antiparallel β-barrel. The three β-hairpins form a cap on one end of the barrel (on the bottom in this figure.)

In comparison, we ran the iterative sequence-based method `PSI-BLAST` [1], the hidden Markov model program `HMMer` [11], and the publicly available threading program `Threader` [14], to see if they could predict any of the known β-trefoils from their sequences (see Section 5). `PSI-BLAST` and `HMMer` [11] both primarily find with reasonable confidence levels sequences from the same family as the query sequence. `Threader` ranked ten or more non-β-trefoils above all trefoils in its library for most members of the non-redundant β-trefoil database. Thus, aligning to a trefoil structural template, and then threading, as we provide in `Wrap-and-Pack`, is clearly warranted.

A server running `Wrap-and-Pack` will be available on the Internet, at `theory.lcs.mit.edu/wrap-and-pack`. This site will also contain a list of high-scoring protein sequences. This paper has also been accepted to the Eighth Annual International Converence on Research in Computational Biology (RECOMB 2004).

Chapter 2

The Algorithm

We begin by describing a β-trefoil template, and then describe the wrapping and packing phases of the algorithm. The `MEMSAT` algorithm [16] is used to filter transmembrane regions before the wrapping phase as the algorithm favors long runs of hydrophobic residues.

2.1 Construction of an abstract structure template

A structural alignment of the β-trefoils in our trefoil database (see Section 3.1) is obtained from the FSSP database [13]. Some of these alignments were hand-curated due to missing atomic coordinate data. The structural alignments were used to deduce an abstract template for the β-trefoil structure, as described below.

A *leaf template* consists of:

1. a B1-strand, followed by a T1 turn of length 2 to 17,
2. followed by a B2-strand, followed by a T2 turn of length 0 to 11, followed by a B3-strand (i.e., the *cap template*),
3. followed by a T3 turn of length 4 to 20, followed by a B4-strand.

A *trefoil template* consists of three leaf templates, separated by T4 turns of length 0 to 16. In addition, the template has a minimum size of 26 and a maximum of 64.

All the β-strands are five residues in length and pleated with the first residue buried. Note that the barrels have a shear of two, where the first three residues of B1 and B4 in the same leaf align, and the last three residues of adjacent B4 and B1 strands in different leaves align. The lengths of the strands were chosen from the residue positions in common in the strands of the structural alignments. The range of turn lengths was automatically determined by the program from the trefoil database by eliminating the two longest and shortest turns, allowing the remaining range of turns extended by length two on each end. In addition, the total length allowed for the leaf template was also upper and lower bounded similarly to the turn ranges.

2.2 Wrapping

We first describe the wrapping phase of the algorithm; the packing phase is below. The wrapping phase of the `Wrap-and-Pack` program is a novel "wrapping" algorithm that searches for the aligning antiparallel β-strands in the cap and the barrel of each leaf, and then in the barrels of neighboring leaves. This phase is somewhat similar to what is done in `BetaWrap`, our previous program to predict right-handed parallel β-helix proteins; however, this phase is much more complicated when applied to the β-trefoils.

2.2.1 First stage: wrapping a cap

The cap wrapping phase attempts to first wrap a subsequence onto the cap template. A score is assigned to each substring that fits to the cap template. The score is computed based on the alignment of the residues of the antiparallel B2 and B3 strands. As the strands are antiparallel, the first residue of B2 is aligned to the last residue of B3, and so on. To score aligned residue pairs in the B2 and B3 strands, a database, called the *twisted β-structure database* (see Section 3.1), was constructed of β-sheets which share with the β-trefoils the properties that the β-sheets are twisted and one face is buried and one exposed (the β-trefoils themselves were excluded from this database to avoid overtraining).

The conditional probability that a residue of type X will align with residue Y, divided by the frequency of residues of type X, given their orientation relative to the core, was estimated from the twisted β-structure database using standard methods (see, e.g. Berger [4]). The natural logarithm of this probability gives the *pair score* of a vertical alignment of two residues. The conditional probability estimates for all the stacking pairs of residues in inward and outward pointing β-sheets learned from the twisted β-structure database, have been reproduced in the Appendix in Tables A.1 and A.2. In addition, conditional probabilities for kitty-corner pairs of residues, i.e. those residues one off from the vertical alignment in either direction, are similarly computed (see Table A.3 in Appendix). Conditional probabilities for inward and outward pointing residues are likewise calculated separately. Either of these two tables can be obtained by flipping the other along the diagonal, so we include only one. All of the cap pairwise alignment scores are modified to be an average of the natural logarithm of the conditional probabilities of residue X aligning with residue Y and residue Y aligning with residue X, since there is not really a direction to the wrap.

For a pair of aligned stands, the *β-sheet alignment score* is the weighted sum of the five alignment scores for the aligned pairs in the β-sheets B2 and B3. A weight of 1 is given to the scores for inward and outward pairs, and 1/2 for the scores of the kitty-corner pairs, to reflect the fact that there are roughly twice as many kitty-corner pairs. Note that one member of the kitty-corner pair is allowed to extend beyond the five-residue β-strand template for a total of 12 such pairs.

The β-sheet alignment score is the heart of the cap recognition method; however, we improve its performance with several bonuses and penalties. The bonuses and penalties are added as real values to the raw alignment scores:

- The gap penalty is given by $|gap - avggap|/stddev$, where *gap* is the length of the turn, *avggap* is the average gap length of that particular type of turn in the allowed turn length range, and *stddev* is the standard deviation of the turn length from the average. The gap penalty in B2-T2-B3 of the cap template is used.

- The β-strand bonus is a bonus of 1 point for every residue in B2 or B3 that the PSIPRED algorithm [15] predicts to be in a β-strand.

We also incorporate several filters into the cap wrapping algorithm:

- The turn filter is applied to T2 turns. Caps that do not pass the filter are discarded. The turn filter looks in the T2 turn region and the last four and first four residues of B2 and B3 for all eight-residue substrings (β-turns are four residues). For each of the four turn tables (see Section 3.1), the logs of the pairwise conditional probabilities of each of the $\binom{8}{1} + \binom{8}{2}$ pairs of residues are added. Caps for which no score for any type of turn is above two are discarded.

- The cap score filter removes caps whose scores fall below a cutoff value of -34.55. This value was determined by computing the scores of the caps of all β-trefoil leaves in the database, eliminating the lowest two scores, and taking the remaining lowest score. The same method was used to pick all score cutoffs.

We remark that while the raw strand-strand score is based on non-β-trefoil stacking preferences, and thus is the same over all cross-validation experiments, the values of these modified cutoffs, bonuses, and penalties varied slightly, based on what β-trefoil structures were included in the training set (see Section 3.2). The numeric values reported for the gap score above, for the gap length ranges above, for the cap and barrel cutoff scores, and for the α-helical penalty later in the manuscript are the ones for the final version of `Wrap-and-Pack` that incorporate all structures in a non-redundant trefoil database.

2.2.2 Wrapping a leaf

The leaf wrapping phase assembles a leaf by integrating caps chosen above with potential B1- and B4-strands. Every position in the sequence is considered the start of a B1-strand, and the top ten scoring caps for it are determined, each uniquelly identifying a B1-T1-cap region. The turn filter described in Section 2.2.1 is applied to the T1-turns, and regions that do not pass the filter are discarded. The score

for a B1-B4 alignment is determined using the β-sheet alignment scores (described above) of their three aligned residues. A β-strand bonus of 1 point for every residue in B4 that PSIPRED [15] predicts to be in a β-strand is added to this score. A global leaf-length penalty, calculated as the gap penalty above for the entire leaf length, is added to the B1-B4 alignment score, producing the barrel score. The B1 PSIPRED predictions are not used here because the B1-B4 alignment score is based on the probability of B4 aligning with B1, assuming the B1 is correct. The barrel scores are filtered against a score cutoff of -24.68, computed similarly to the cap score filter.

The score of a leaf is the cap score divided by 2.75, plus the barrel score, plus two gap penalties for T1 and T3, calculated as above. The divisor is roughly based on the ratio of the cap to barrel scores. The top five scoring leaves for each B1-strand are stored for use in wrapping entire trefoils.

From a leaf to multiple leaves

Beginning with each potential leaf from the previous section, the five top scoring aligned leaves are calculated forward in sequence as follows: From potential leaves that are within a gap length of 0 to 16, the score for a B4-B1 alignment from the first to the second leaf is determined using the β-sheet alignment scores (described above) of their three aligned residues; a gap penalty and the scores for both leaves are added to the score. The β-strand bonus of 1 point per residue in B1 that PSIPRED [15] predicts to be in a β-strand is also added. This process is repeated with the third leaf, producing a search tree of degree five and depth two, where the tree leaves correspond to potential trefoils.

A final trefoil score is assigned to each of these trefoils by incorporating the β-sheet alignment scores for the B4-B1 alignment from the third trefoil leaf to the first. If any of the 3 B4-B1 alignment scores falls below a cutoff of -24.18, determined similarly to the cap and barrel score filters, the trefoil is discarded. Then an alpha penalty is applied to this score to get rid of those potential trefoils with too much α-helical content. The number of putative β-strands, other than the B4-strand, that have α-content is determined by PSIPRED [15]. If there are more than two such strands in an

entire trefoil, an empirical penalty of five times the number of additional β-strands with α-helical content, is applied. Note that in the actual trefoils the B4-strand is often directly preceded by an α-helix, so it is left out of the alpha penalty.

We also incorporate several filters to the assembled trefoils. If the size of the entire trefoil falls outside the allowed range of the trefoil template, the putative trefoil is eliminated. Based on the observation that the sizes of the gaps between cap strands are roughly the same in a given trefoil, the trefoils that do not meet these criteria are filtered out: If the difference between the last two cap gaps is one, or the difference between any of the gaps is strictly greater than three, that putative trefoil is eliminated.

The remaining trefoils with the top four final scores are retained.

2.3 Packing

With only wrapping and not packing, preliminary results indicated that we achieve accurate strand prediction of the known β-trefoils, however, we have a high rate of false positives. To filter out non-β-trefoils, we used a backbone dependent rotamer library from the `SCWRL` program [21] to find an energy-favorable placement of the sidechains, at least those that pack within the interior of the β-trefoil.

For each of the top four trefoils from above, each pair of aligning cap strands is run through `SCWRL` to get an energy score from threading onto leaf 2 of all six representative structures in the superfamily database. (Leaf 2 was chosen because it was the most conserved among the structural alignments of the trefoils.) The total energy score for a trefoil against each of the representative structures is the sum of the energy scores for all three of its cap strands. The lowest total energy score of the six is the energy score for that putative trefoil. These trefoils are filtered against an energy score cutoff of 15, determined similarly to the turn ranges in Section 2.1. Only the cap strands are used because `SCWRL` has trouble with side-chain placement in the tightly-packed barrel.

The *sequence score* is the score of the top scoring trefoil that passed the filter.

Chapter 3

Methods

3.1 The databases

The PDB-minus database was constructed from the NCBI non-redundant protein structure database (P-value of 10^{-80}; 04-03 update), with the β-trefoils removed. The database has 6996 structures.

The twisted β-structure database was created from PDB-minus by looking for alternating patterns of residue accessibility in twisted β-strands. The PDB-minus structure files were processed using the program `Stride` [12], which annotates secondary structure, hydrogen bonds, and residue accessibilities. Twisted β-strands were determined by considering whether or not at least two-thirds of the angles between every other CA-CB vector in the strand have a cosine value of less than 0.8. Note that this measure is intended to capture the twisted nature of β-sheets like those of the trefoils, which contrast with the very straight topological β-sheets of the β-helices. In all, 2656 protein chains from PDB-minus contributed sheets or portions of sheets to the database. Tables A.1, A.2, and A.3 in the Appendix present the scores for the pairwise amino acid stacking and kitty corner probabilities learned from this database.

The four turn databases were comprised of β-turns from each of Types I, I', II, and II' (according to the `Stride` program) in PDB-minus. Only β-turns between two β-strands with a maximum gap length of 20 were added to the database. The four

turn conditional probability tables were constructed by taking the frequency of all pairs of residues and individual residues in each pair of positions in the turn region and dividing by the number of such turns.

The β-trefoil database was taken from the NCBI non-redundant protein structure database (P-value of 10^{-80}; 04-03 update). The database contained 25 structures, which come from ten families of closely related proteins in the SCOP [18] database. The superfamily database contained one representative trefoil structure from each of the six SCOP superfamilies chosen from the β-trefoil database. Potential new β-trefoils were identified from the sequence database SWISS-PROT [2].

3.2 Training

A ten-fold cross-validation was performed on the ten β-trefoil families of closely related proteins in the SCOP [18] database. For each cross, proteins in one β-trefoil family were placed in the test set, while the remainder of the β-trefoils were placed in the training set. The scores reported for the β-trefoil proteins in Table 3.1 and in Figure 4-1 are the scores in the leave-family-out cross experiment for that β-trefoil's protein family. The optimal thresholds for the gap score, the gap length ranges, the cap and barrel cutoff scores, and the α-helical penalty (as described in Section 2.2.1 above), were optimized for training data, and thus recalculated for each experiment. Note that the PDB-minus database was not used for setting any of these values; it was only used to generate the conditional probability tables.

SCOP Family	Name	Source	PDB	Score	P-value
Fibroblast growth factors	Basic FGF (FGF2)	*Homo sapiens*	1BFF	−126.36	0.054
Fibroblast growth factors	Acidic FGF (FGF1)	*Homo sapiens*	2AFG	−121.44	0.023
Fibroblast growth factors	Acidic FGF (FGF1)	*Notophthalmus viridescens*	1FMM	−124.01	0.038
Fibroblast growth factors	FGF4	*Homo sapiens*	1IJT	−124.37	0.040
Fibroblast growth factors	Keratinocyte GF (FGF7)	*Rattus norvegicus*	1QQK	−125.56	0.048
Fibroblast growth factors	Keratinocyte GF (FGF7)	*Rattus norvegicus*	1QQL	−120.51	0.018
Fibroblast growth factors	FGF9	*Homo sapiens*	1G82	−126.64	0.057
Fibroblast growth factors	FGF10	*Homo sapiens*	1NUN	−128.44	0.077
Interleukin-1	Interleukin-1β	*Homo sapiens*	2I1B	−123.09	0.033
Interleukin-1	Interleukin-1β	*Mus musculus*	2MIB	−120.65	0.018
Interleukin-1	I-1 receptor antagonist	*Homo sapiens*	1IRP	−119.25	0.013
Interleukin-1	Interleukin-1α	*Homo sapiens*	2ILA	−121.44	0.023
IP3 receptor type 1	IP3 receptor binding core	*Mus musculus*	1N4K	−127.57	0.066
Ricin B-like	Plant cytotoxin B-chain (lectin)	*Abrus precatorius*	1ABR	−123.94	0.037
Ricin B-like	Plant cytotoxin B-chain (lectin)	*Sambucus ebulus*	1HWM	−122.49	0.029
Cysteine rich domain	Mannose receptor	*Mus musculus*	1DQG	−144.13	0.277
Agglutinin	Agglutinin	*Amaranthus caudatus*	1JLX	−122.59	0.027
Kunitz (STI) inhibitors	Winged bean albumin 1	*Psophocarpus tetragonolobus*	1WBA	−124.00	0.038
Kunitz (STI) inhibitors	Trypsin inhibitor	*Erythrina caffra*	1TIE	−124.04	0.038
Kunitz (STI) inhibitors	chymotrypsin inhibitor WCI	*Psophocarpus tetragonolobus*	2WBC	−122.64	0.030
Kunitz (STI) inhibitors	Soybean trypsin inhibitor	*Glycine max*	1BA7	−124.99	0.044
Kunitz (STI) inhibitors	Amylase/subtilisin inhibitor	*Hordeum vulgare*	1AVA	−127.77	0.069
Clostridium neurotoxins	Tetanus neurotoxin	*Clostridium tetani*	1A8D	−119.25	0.013
Fascin	Fascin	*Homo sapiens*	1DFC	−120.30	0.017
HIS-rich actin-binding	Hisactophilin	*Dictyostelium discoideum*	1HCD	N/A	N/A

Table 3.1: The sequences in the β-trefoil database and their scores. SCOP families are separated by single lines and superfamilies by double lines. Note that 1HCD had no potential trefoil pass the filters.

Chapter 4

Results

`Wrap-and-Pack` achieves 92% specificity and 92.3% sensitivity for a score of -128.44 when run on the NCBI non-redundant protein structure database (Figure 4-1). The score for each β-trefoil is taken from its cross-validation run. In Table 3.1, the β-trefoil proteins used in this study are listed along with their cross-validation scores. The P-value of a sequence is given by the portion of the PDB-minus that higher than the sequence.

The top five non-β-trefoils proteins are a surface layer protein (1L0Q) (A seven-stranded β-propeller) from the Archaeon *Methanosarcina mazei*, with a score of -105.1; the Growth-arrest-specific protein Gas6 (1H30) (Laminin G-Domain Protein) from *Homo sapiens*, with a score of -107.3; β-Galactosidase (1BGL) (A Hydrolase) from *Escherichia coli*, with a score of -110.5; the Tombusvirus coat protein (2TBV) (An all beta protein) from Tomato bushy stunt virus, with a score of -110.7; and chain B of Centromere DNA-Binding Protein Complex Cbf3 Subunit D (1NEX:B) (A 7-bladed beta-propeller WD40-repeat) from *Saccharomyces cerevisiae*, with a score of -111.2. The other high-scoring non-β-trefoils tended to be all beta and beta/alpha proteins with seven or more β-strands located within a contiguous section of less than 200 residues.

The `Wrap-and-Pack` program has identified many new sequences that may contain β-trefoils. Table 4.1 lists some examples of the predicted proteins from SWISS-PROT [2]. A number of these may be related to the known β-trefoils. The human counterpart

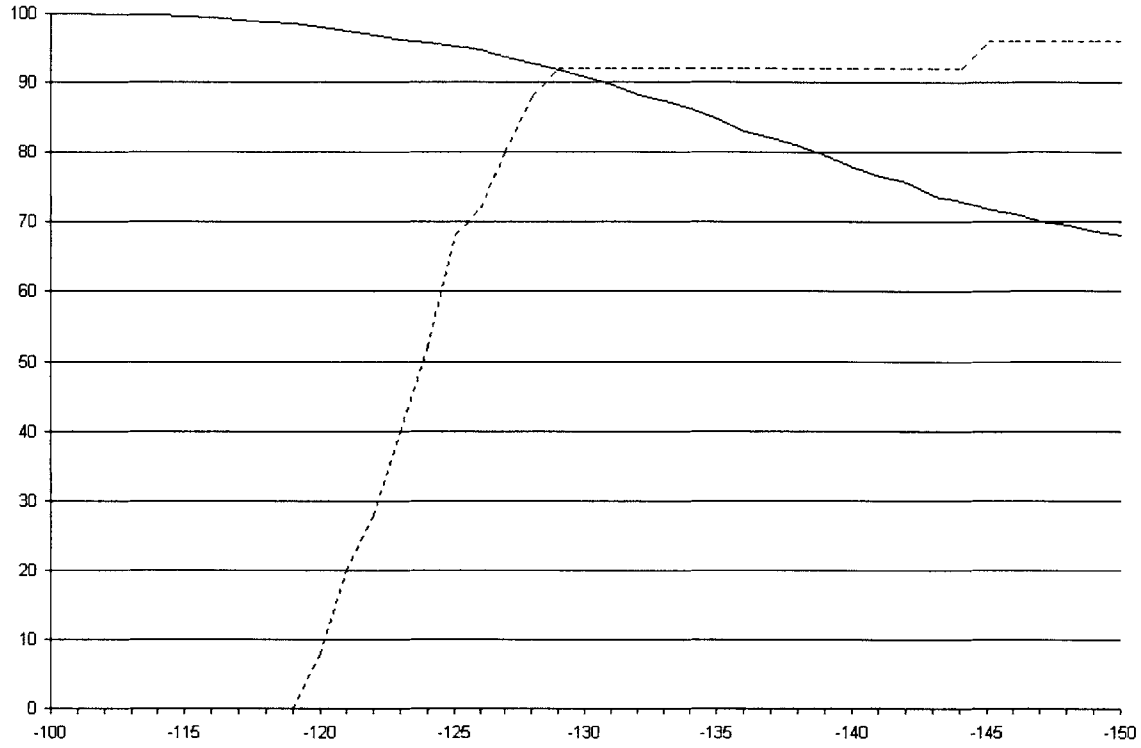

Figure 4-1: Percent sensitivity and specificity as a function of β-trefoil protein score cutoff, as computed by `Wrap-and-Pack`. Sensitivity is represented as a solid line, and specificity as a dashed line. The NCBI non-redundant protein structure database (NR), which was used to compute these values, contains 25 β-trefoils and 6,996 non-β-trefoils. The score for each β-trefoil is taken from its cross-validation run.

to the african clawed frog fascin protein is known to be a β-trefoil. The lectin precursor has related proteins that are known to be β-trefoils; however, most certainly all related proteins with solved structures are not β-trefoils.

A more complete list will be maintained at `theory.lcs.mit.edu/wrap-and-pack`.

ID	Description	Organism	Score
Q92176	Coronin-like protein p57	*Bos taurus*	−104.15
O13798	Hypothetical protein	*Schizosaccharomyces pombe*	−107.02
P24821	Tenascin [Precursor]	*Homo sapiens*	−107.57
P20930	Filaggrin [Precursor]	*Homo sapiens*	−108.78
Q9UBF2	Coatomer gamma-2	*Homo sapiens*	−109.14
P33194	DNA damage binding	*Homo sapiens*	−110.28
Q91837	Fascin	*Xenopus laevis*	−113.22
Q01806	Lectin 1 [Precursor]	*Medicago truncatula*	−115.12

Table 4.1: Examples of proteins predicted to form β-trefoils and their scores. Identifiers are taken from SWISS-PROT.

Chapter 5

Comparison with Other Methods

We tried three existing computational methods to see how they performed in terms of their ability to detect the relationships between the known families of β-trefoils: `PSI-BLAST` [1], `HMMer` [11], and `Threader` [14].

First the 25 sequences in the β-trefoil database were used to search the NCBI nonredundant database (27-Aug-2003 update, 1,486,004 entries) using the iterative multiple sequence alignment program `PSI-BLAST` [1] (version 2.2.2). The default E-value cutoff for inclusion of 0.001 was used; all searches converged before 10 rounds (see Table 5.1). A sequence was considered as having been found if it was included in the profile after any of the rounds. Only 14 of the 25 sequences gave profiles that included sequences from other β-trefoil families. Of those, 10 included only a single sequence from another family. Running `PSI-BLAST` against a larger database, such as SwissProt [2], reduces E-values so that matches across familes occur more infrequently.

Next, we looked at the `HMMer` [11] hidden Markov model program. The input to `HMMer` is a multiple sequence alignment (MSA). Pfam [3] is a database of protein families created by `HMMer` from hand-curated multiple alignments. Table 5 contains a list of all the families that contain any of the 25 sequences in the β-trefoil database. As these were created by `HMMer`, they give an idea of the algorithm's performance of the β-trefoils. Only one Pfam family, PF00652, has members from two different SCOP families.

	1BFF	2AFG	1FMM	1IJT	1QQK	1QQL	1G82	1NUN	2I1B	2MIB	1IRP	2ILA	1N4K	1ABR	1HWM	1DQG	1JLX	1WBA	1TIE	2WBC	1BA7	1AVA	1A8D	1DFC	1HCD
1BFF	X	X	X	X	X	X	X	X	X	X														X	
2AFG	X	X	X	X	X	X	X	X																	
1FMM	X	X	X	X	X	X	X	X									X								
1IJT	X	X	X	X	X	X	X	X																	
1QQK	X	X	X	X	X	X	X	X																	X
1QQL	X	X	X	X	X	X	X	X																	
1G82	X	X	X	X	X	X	X	X																	
1NUN	X	X	X	X	X	X	X	X																	
2I1B	X	X	X	X	X	X	X	X	X	X	X	X													
2MIB	X								X	X	X	X													
1IRP	X								X	X	X	X													
2ILA									X	X	X	X													
1N4K													X												
1ABR														X	X	X									
1HWM														X	X										
1DQG														X		X									
1JLX			X														X								
1WBA																		X	X	X	X	X	X		
1TIE																		X	X	X	X	X			
2WBC													X					X	X	X	X	X			
1BA7																		X	X	X	X	X	X		
1AVA																		X	X	X	X	X			
1A8D																							X		
1DFC	X																							X	X
1HCD					X																			X	X

Table 5.1: Results of PSI-BLAST searches on the known β-trefoil structures. An 'X' indicates that the protein in that column, indexed by its PDB code (see Table 1) was found when PSI-BLAST was run on the protein indexing the given row. SCOP families are separated by single lines and superfamilies by double lines.

Pfam	PDB codes
PF00167	1BFF 1G82 1IJT 1QQK 1QQL 2AFG
PF00047	1NUN
PF02394	2I1B 2ILA 2MIB
PF00340	1IRP 2I1B 2ILA 2MIB
PF01365	1N4K
PF00161	1ABR
PF00652	1DQG 1HWM
PF07468	1JLX
PF00197	1BA7 1TIE 1WBA 2WBC
PF00128	1AVA
PF01742	1A8D
PF06268	1DFC
PF06402	1HCD

Table 5.2: List of all the Pfam families that contain any of the 25 sequences of the β-trefoil database and the β-trefoils they contain. Single lines separate SCOP familes and double lines separate superfamilies. Note that PF00652 is the only Pfam family that contains sequences from multiple SCOP families.

Next we ran `HMMer` 2.3.1 seeded with structural alignments from the standard structural alignment database FSSP [13]. We remark that, because the FSSP MSAs are typically constructed using full information as to how the structures align in 3D, this is not a strictly sequence-based method. In spite of this advantage, we find it performs poorly in the prediction of β-trefoils. There are 12 β-trefoil structures in FSSP. Each of these structures was used as a seed to create a multiple sequence alignment of all the trefoils in FSSP (see Table 5.3). Each of these initial alignments was then used to do a separate leave-one-out cross validation. One sequence is removed from the alignment, which is then used to create an HMM with `HMMer`. The HMM is then used in an attempt to identify the missing sequence. Note that a dash in the table indicates that the protein left out of the cross validation is the seed for the FSSP alignment. Thus Table 5.3 theoretically represents 12 by 11 cross validation experiments; however, since 1JLY and 3BTA have no FSSP alignment with each other, the table is not exactly 12 by 11 in size. As can be seen, each one of the 12 seeded HMM models failed to find any other β-trefoils in the table, except for

Human Interleukin-1β (2I1B), which was found by the alignment seeded by Human Interleukin-1 receptor antagonist protein (1IRP), which are both in the same family. Thus, this corresponds to sensitivity less than 20%.

Finally, the program `Threader 3.4` [14] was used to thread the 25 sequences in the β-trefoil database onto an accompanying fold library (2-03 version, 5335 domains). Threadings were sorted by the Z-scores given by the program, where a high score indicates that a protein is a good match to the target structure. For each trefoil, domains of all trefoils in the family were removed from the database. Table 5.4 gives the number of non-β-trefoils in the `Threader` fold library that score above any β-trefoil from a different family in the library. Only two (1DQG and 1WBA) of the trefoils were matched to trefoil structures with highest confidence. Fifteen of the trefoils were matched to ten or more non-β-trefoils with higher confidence than any trefoil.

`Threader` was also used to thread all sequences in the non-redundant PDB onto all β-trefoil structures its database (see Figure 5-1). Only β-trefoil structures were used to reduce computation time. The weighted sum `Threader` reports of its two energy functions was used to create a graph of specificity and sensitivity as a function of cutoff score. Only the score of the lowest scoring (best matching) trefoil structure was reported for each sequence. For the actual β-trefoil sequences, the lowest scoring β-trefoil structure from another family was used. As can be seen by comparing Figures 4-1 and 5-1, `Wrap-and-Pack` outperforms `Threader` in this test.

	1BFG	2I1B	1IRP	2ILA	1ABR	1DQG	1JLY	1WBA	1AVA	1A8D	1DFC	1HCE
1BFG	-											
2I1B		-										
1IRP		X	-									
2ILA				-								
1ABR					-							
1DQG						-						
1JLY							-				-	
1WBA								-				
1AVA									-			
1A8D										-		
1DFC							-				-	
1HCE												-

Table 5.3: Results of `HMMer` searches with the rows representing the initial seed aligned by FSSP (all β-trefoils in FSSP were used as seeds), and the columns, these structures, indexed by their PDB codes (see Table 1). An 'X' indicates that the protein in that column was found when it was left out of the multiple sequence alignment fed to `HMMer`. A dash indicates a protein is the seed. FSSP does not contain alignments for 1JLY and 1DFC with each other. Single lines separate SCOP families and double lines, superfamilies.

1BFF	2AFG	1FMM	1IJT	1QQK	1QQL	1G82	1NUN
8	5	4	63	12	7	37	56

2I1B	2MIB	1IRP	2ILA	1N4K	1ABR	1HWM	1DQG
87	61	1	2	62	77	10	0

1JLX	1WBA	1TIE	2WBC	1BA7	1AVA	1A8D	1DFC	1HCD
40	0	4	25	119	2	62	73	17

Table 5.4: Results of **Threader** searches. The table, split into three rows, represents a β-trefoil threaded onto all structures in the **Threader** library; the values represent the number of non-β-trefoil structures in the library which had a higher Z-score than the highest scoring trefoil from a different family. The double lines indicate superfamily divisions, while the single lines, family.

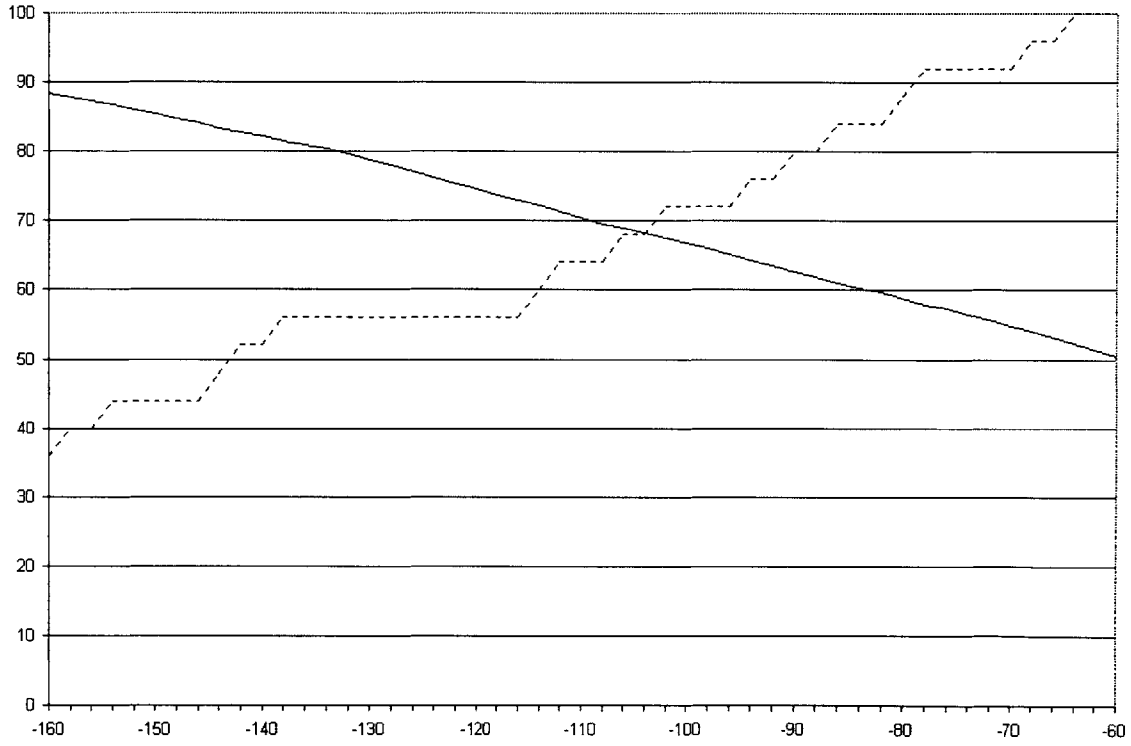

Figure 5-1: Sensitivity and specificity of β-trefoil prediction of Threader when run using only the β-trefoil structures as a function of raw score cutoff. Sensitivity is represented as a solid line, and specificity as a dashed line. The NCBI non-redundant protein structure database, which was used to compute these values, contains 25 β-trefoils and 6,996 non-β-trefoils. The score for each β-trefoil is taken from its cross-validation run.

Chapter 6

Discussion

Our results indicate that the methods that we developed for `BetaWrap` can be generalized to other primarily beta folds, and in particular, the β-trefoil. These results are achieved by modifying the wrapping algorithm to reflect a different strand topology and turn distribution and replacing bonuses particular to β-helices with a set learned from the trefoils. The wrap is used as a guide for the packing, or threading, component of the algorithm, where 3D energetic information is incorporated into our structural template to filter out false positives. The `Wrap-and-Pack` algorithm, introduced here, is able to identify β-trefoils with 92% specificity and 92.3% sensitivity.

We are pursuing a number of directions to improve these results. For the packing phase of the algorithm, instead of `SCWRL` we intend to use `Threader`, which takes into account solvent accessibility in addition to atom collisions. This could be accomplished by modifying `Threader`'s database to contain only structures of β-trefoil domains. Then we need only compare a sequence that matches our structural template against the modified database. In addition, `Wrap-and-Pack` tends to find certain other classes of close-packed proteins (see Section 4); thus a logical step is to filter out obvious false positives, such as the WD40 β-propellers. Such a filter would certainly help in combination with an iterative bootstrapping procedure, whereby newly identified sequences are incorporated into the training set and aid in the identification of more distant families; see for example [5]). Unfortunately, preliminary attempts with comparative genomics did not improve `Wrap-and-Pack`'s performance.

We plan to apply the methods described here to other mainly β-folds, such as the β-propellers, some of which score highly with `Wrap-and-Pack` for trefoils, β-rolls, and β-clams. We hope to further automate our semi-automatic algorithm construction. Currently, the bonuses and penalties were set based on performance on the known trefoils (leaving out trefoils in the same family in cross-validation, to avoid overtraining), but they were set by hand: a general method that could automatically learn which features of the known structures of a fold template contributed most to seperation of positive and negative examples of the fold could speed up development of the wrapping phase for new sequence-heterogeneous beta-structural motifs. We will similarly use structural alignments from FSSP to construct abstract structural templates for these folds. We intend to use the `Trilogy` program of Bradley, Kim, and Berger [8] to search for sequence-structure patterns within these folds to incorporate accurate bonuses and penalties. `Trilogy` has already identified a sequence-structure pattern which occurs in a single blade of β-propeller proteins in different superfamilies of the β-propeller fold [8]. `Trilogy` has also been extended to look for patterns within superfamilies of the same fold.

Appendix A

Pairwise Alignment Probability Tables

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	5.8	4.3	10.0	8.3	5.8	5.2	5.4	6.7	5.9	8.0	7.4	5.9	6.1	6.4	5.4	5.5	7.9	8.6	4.9	4.4
C	2.4	18.4	2.9	3.3	2.8	4.7	4.7	2.5	3.3	2.5	2.3	7.1	4.6	3.7	3.7	3.3	2.7	2.4	15.6	3.9
D	3.2	1.7	3.3	2.2	1.4	1.3	4.7	1.9	3.2	1.1	2.0	5.3	1.8	1.6	3.7	2.7	2.7	2.6	2.4	1.1
E	3.2	2.3	2.7	2.1	2.3	1.3	2.9	2.0	4.0	2.1	2.5	5.0	2.5	4.4	5.9	3.3	4.0	2.0	3.8	2.8
F	6.2	5.4	4.8	6.4	9.3	9.5	5.8	8.9	5.3	6.4	10.1	3.4	6.1	6.0	7.4	6.0	5.3	7.7	8.2	8.4
G	3.2	5.3	2.5	2.1	5.5	6.5	4.7	4.2	1.7	4.0	3.2	4.3	6.4	3.3	2.2	4.8	4.0	4.8	4.5	4.2
H	1.0	1.6	2.7	1.4	1.0	1.4	2.2	1.0	1.0	1.1	1.8	1.5	2.1	1.5	1.9	1.3	2.1	0.8	0.4	1.8
I	12.1	8.1	10.4	9.5	15.0	12.3	9.8	17.4	11.5	14.2	7.9	5.9	6.4	13.0	9.5	9.2	9.2	12.2	10.7	10.3
K	2.7	2.7	4.6	4.8	2.3	1.3	2.5	3.0	2.9	1.9	3.1	3.1	1.1	4.8	2.2	4.5	4.3	3.4	5.3	6.2
L	15.1	8.4	6.5	10.0	11.2	12.1	10.9	14.8	7.5	17.0	12.7	10.8	11.1	12.3	9.8	10.9	9.9	14.7	9.1	9.7
M	2.7	1.6	2.3	2.4	3.5	1.9	3.6	1.6	2.4	2.5	6.1	3.7	5.7	1.5	4.0	1.9	3.5	2.1	1.6	2.4
N	1.3	2.7	3.5	2.8	0.7	1.5	1.8	0.7	1.4	1.3	2.2	4.3	1.8	2.6	2.1	1.6	2.9	0.8	1.1	1.5
P	1.1	1.6	1.0	1.2	1.1	1.9	2.2	0.7	0.4	1.1	2.9	1.5	2.9	0.7	0.8	1.3	1.4	1.0	3.8	2.2
Q	2.3	2.4	1.9	4.1	2.1	1.9	2.9	2.6	3.7	2.4	1.4	4.3	1.4	1.5	2.4	3.0	1.6	2.1	1.1	5.1
R	2.3	2.7	4.8	6.4	2.9	1.5	4.3	2.2	2.0	2.2	4.5	4.0	1.8	2.7	2.6	2.5	4.2	2.2	3.3	5.5
S	2.5	2.6	3.8	3.8	2.5	3.4	3.3	2.3	4.3	2.6	2.3	3.4	3.2	3.7	2.7	6.3	4.4	3.1	3.3	3.0
T	5.2	3.2	5.6	6.9	3.3	4.3	7.6	3.4	6.2	3.5	6.3	9.0	5.0	2.9	6.7	6.5	6.8	3.8	1.6	6.0
V	22.4	11.1	21.5	13.4	18.7	20.2	11.6	17.7	18.9	20.5	14.7	9.9	13.9	15.2	13.8	17.7	14.8	19.8	10.2	15.1
W	1.5	8.4	2.3	2.9	2.3	2.2	0.7	1.8	3.4	1.5	1.3	1.5	6.1	0.9	2.4	2.2	0.7	1.2	2.2	2.4
Y	3.5	5.6	2.7	5.9	6.3	5.5	8.0	4.6	10.8	4.2	5.2	5.6	9.6	11.2	10.6	5.4	7.3	4.7	6.5	3.7

Table A.1: Conditional probabilities for alignment of buried residues from the twisted β-structure database. The value in row i, column j is 100*P(seeing residue i, given that it is aligned with residue j).

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	4.8	4.9	4.9	2.8	5.5	7.2	7.2	4.7	4.0	5.7	3.0	4.2	3.9	5.7	3.1	3.3	2.8	5.1	4.0	4.5
C	2.1	16.5	2.5	1.0	1.9	2.5	0.9	2.6	0.9	1.7	3.0	0.7	1.1	1.3	1.5	1.2	1.1	2.0	1.5	2.0
D	4.0	4.6	3.9	2.7	2.7	4.1	4.5	2.7	6.1	2.1	3.0	4.2	3.7	4.1	6.0	4.2	4.6	2.3	2.5	2.1
E	3.9	3.2	4.7	3.1	4.6	3.1	6.9	6.1	12.1	5.0	3.0	8.1	6.0	5.2	11.1	6.3	9.8	5.7	5.5	4.5
F	5.8	4.6	3.5	3.4	6.8	9.3	3.7	4.7	3.5	4.3	4.2	2.7	6.9	6.5	4.5	3.7	2.4	5.3	5.0	6.6
G	5.8	4.6	4.0	1.8	7.1	6.5	3.5	3.8	2.3	3.2	5.6	2.9	3.0	3.1	1.8	2.8	2.7	4.4	2.7	3.6
H	4.1	1.2	3.1	2.8	2.0	2.5	3.7	1.7	1.9	1.7	1.6	2.5	3.4	2.2	2.3	4.3	3.1	2.1	3.0	2.5
I	7.6	10.0	5.5	7.0	7.2	7.7	5.0	14.2	6.2	9.1	7.5	3.0	5.7	5.0	6.4	5.1	3.9	9.0	8.7	6.8
K	6.2	3.2	11.7	13.3	5.2	4.3	5.2	5.9	4.5	6.9	11.0	6.4	3.7	6.2	3.4	8.3	9.5	6.3	6.5	6.8
L	10.7	7.5	4.9	6.8	7.7	7.6	5.8	10.7	8.5	14.5	10.7	6.4	8.0	7.3	8.7	6.4	4.8	10.1	10.0	7.3
M	1.4	3.2	1.7	1.0	1.8	3.1	1.3	2.1	3.2	2.5	1.9	1.7	3.9	2.2	1.1	2.1	1.5	1.5	3.2	2.4
N	2.6	1.0	3.3	3.6	1.6	2.2	2.8	1.2	2.6	2.1	2.3	7.4	2.5	4.3	2.3	3.8	4.0	2.1	2.7	2.9
P	1.8	1.2	2.1	1.9	3.0	1.7	2.8	1.6	1.1	1.9	4.0	1.9	4.1	1.7	2.1	1.9	2.0	1.7	7.7	1.7
Q	5.5	2.9	4.8	3.5	5.9	3.7	3.7	2.9	3.8	3.7	4.7	6.6	3.4	5.7	4.3	5.2	5.7	3.2	3.7	3.7
R	4.3	4.6	10.3	10.9	5.9	3.1	5.6	5.4	3.1	6.3	3.5	5.1	6.2	6.2	3.4	6.3	6.5	8.2	7.7	6.8
S	4.2	3.6	6.6	5.8	4.5	4.4	9.8	4.0	6.9	4.3	6.1	7.8	5.3	7.1	5.9	8.8	8.7	3.7	4.2	4.9
T	5.3	4.6	10.8	13.1	4.2	6.4	10.4	4.5	11.6	4.8	6.3	12.1	8.0	11.3	8.9	12.7	13.8	6.5	3.7	5.5
V	12.1	10.7	6.6	9.5	11.9	12.8	8.7	13.1	9.7	12.5	7.7	7.9	8.7	7.9	14.0	6.9	8.2	13.0	8.7	10.7
W	1.7	1.5	1.3	1.6	2.0	1.4	2.2	2.3	1.8	2.2	3.0	1.9	7.1	1.7	2.4	1.4	0.8	1.6	2.5	1.9
Y	6.2	6.3	3.6	4.4	8.7	6.1	6.1	5.8	6.1	5.3	7.5	6.4	5.0	5.3	6.8	5.3	4.0	6.3	6.2	12.6

Table A.2: Conditional probabilities for alignment of exposed residues from the twisted β-structure database. The value in row i, column j is 100*P(seeing residue i, given that it is aligned with residue j).

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	4.9	4.8	4.1	3.5	4.5	5.7	7.8	4.5	2.9	4.4	4.3	4.1	4.4	3.8	3.2	5.2	4.7	4.6	6.7	3.8
C	1.6	2.3	3.1	2.4	1.9	1.6	2.4	1.5	2.6	1.3	1.5	1.5	1.3	1.6	2.6	3.6	1.9	1.7	2.7	2.2
D	3.6	3.9	2.2	2.0	4.1	3.9	3.1	3.4	3.1	3.6	3.7	5.6	4.0	2.1	2.5	3.4	3.3	3.4	2.2	2.8
E	5.5	5.5	5.7	4.3	6.5	6.9	4.5	6.8	6.1	6.8	7.5	4.1	4.2	4.7	5.7	4.7	5.4	7.2	5.5	5.6
F	4.6	4.4	4.2	5.5	4.1	5.0	6.7	5.8	3.3	4.6	5.1	3.0	4.2	4.3	6.0	5.4	4.7	4.2	5.7	4.4
G	5.0	3.9	4.0	5.6	3.2	6.1	4.5	3.5	4.6	4.0	1.7	4.4	5.0	5.3	4.8	5.2	5.5	3.8	4.0	5.3
H	2.9	3.2	1.2	1.8	3.3	2.1	2.0	2.3	1.8	2.8	2.5	1.8	2.5	2.1	2.0	3.0	2.8	2.5	2.6	1.6
I	8.1	5.9	9.0	9.8	6.9	7.3	10.5	7.6	8.2	6.1	6.9	7.7	7.5	6.1	7.9	6.9	9.6	6.9	6.8	6.5
K	6.6	7.7	6.8	6.3	6.9	6.1	5.8	7.3	8.3	6.8	5.7	7.7	5.8	8.2	6.1	3.9	5.0	6.6	5.6	8.2
L	9.6	5.2	11.3	9.6	6.8	8.8	7.6	8.8	7.9	8.9	8.0	7.9	8.5	7.4	10.1	11.2	9.7	8.7	6.3	7.0
M	2.3	1.4	1.7	3.1	1.6	2.1	2.4	2.0	1.9	1.9	2.3	2.0	3.7	1.6	2.1	1.3	2.7	1.7	1.7	1.9
N	3.1	4.8	2.7	2.2	3.4	3.5	3.6	2.9	2.0	2.5	4.0	2.0	2.3	5.0	3.2	3.7	2.2	3.0	1.4	2.7
P	2.8	2.0	2.4	2.3	1.8	1.6	1.6	1.8	2.4	2.0	2.6	2.1	2.9	3.2	2.1	1.9	2.0	2.5	3.1	2.0
Q	4.6	3.2	3.8	2.3	4.2	3.6	3.8	4.4	4.1	4.5	4.3	4.3	2.3	2.2	3.5	2.7	3.7	4.1	6.4	5.5
R	5.8	5.0	6.8	6.1	7.1	5.1	5.1	5.7	6.2	6.4	4.8	7.2	6.2	7.6	5.4	5.0	4.4	6.4	5.6	5.9
S	5.3	7.5	5.2	5.7	7.5	6.8	4.7	6.7	3.4	6.4	4.4	6.3	5.2	6.8	4.1	4.1	3.4	5.1	5.1	6.1
T	7.4	9.6	6.7	5.7	9.7	8.0	6.0	7.9	9.5	9.0	8.8	6.4	6.7	7.4	6.6	8.4	8.3	9.5	6.3	9.0
V	8.9	9.1	13.3	13.1	9.9	8.8	10.7	9.5	9.7	11.0	13.2	11.2	11.8	11.1	11.2	11.4	11.7	11.4	7.7	9.8
W	1.4	1.7	1.9	3.0	1.8	1.1	2.7	1.9	2.3	2.0	1.3	3.0	1.3	1.6	3.2	1.4	2.2	1.5	1.3	1.2
Y	6.0	8.9	3.7	5.9	4.8	5.8	4.5	5.8	9.6	5.0	7.3	7.6	9.8	7.9	7.7	7.5	6.6	5.1	13.1	8.4

Table A.3: **Table A3:** Conditional probabilities for kitty-corner pairs of residues, i.e. those residues one off from the vertical alignment in either direction, from the twisted beta-structure database. Conditional probabilities for inward and outward pointing residues are calculated separately. The value in row i, column j is 100*P(seeing buried residue i, given that it is aligned with a residue adjacent to exposed residue j). Either of these two tables can be obtained by flipping the other along the diagonal, so only one is included.

Bibliography

[1] 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.

[2] A. Bairoch and R. Apweiler. The SWISS-PROT protein database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, 28:45–48, 2000.

[3] A. Bateman, L. Coin, R. Durbin, R. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. Sonnhammer, D. Studholme, C. Yeats, and S. Eddy. A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry*, 33:3038–3049, 1994.

[4] Bonnie Berger. Algorithms for protein structural motif recognition. *J. of Computational Biology*, 2:125–138, 1995.

[5] Bonnie Berger and Mona Singh. An iterative method for improved protein structural motif recognition. *J. of Computational Biology*, 4(3):261–273, Fall 1997.

[6] 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. National Academy of Sciences. USA*, 98(26):14819–14824, 2001.

[7] P. Bradley, L. Cowen, M. Menke, J. King, and B. Berger. Predicting the beta helix fold from protein sequence data. In *Proceedings of the Fifth International Conference on Computational Molecular Biology (RECOMB)*, pages 58–66, April 2001.

[8] Philip Bradley, Peter S. Kim, and Bonnie Berger. Trilogy: Discovery of sequence-structure patterns across diverse proteins. *Proc. National Academy of Sciences. USA*, 99:8500–8505, 2002.

[9] Stephen H. Bryant and Charles E. Lawrence. An empirical energy function for threading protein sequence through the folding motif. *Proteins: Structure, Function and Genetics*, 16:92–112, 1993.

[10] L. Cowen, P. Bradley, M. Menke, J. King, and B. Berger. Predicting the beta-helix fold from protein sequence data. *J. of Computational Biology*, 9(2):261–276, 2002.

[11] S.R. Eddy. Profile hidden Markov models. *Bioinformatics*, 14:755–763, 1998.

[12] D. Frishman and P. Argos. Knowledge-based secondary structure assignment. *Proteins: structure, function and genetics*, pages 556–579, 1995.

[13] L. Holm and C. Sander. Mapping the protein universe. *Science*, 260:595–602, 1996.

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

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

[16] D.T. Jones, W.R. Taylor, and J.M. Thornton. A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry*, 33:3038–3049, 1994.

[17] J. Moult, K. Fidelis, A. Zemla, and T. Hubbard. Critical assessment of methods of protein structure prediction (CASP)-round V. *Proteins: Structure, Function, and Genetics*, 53:334–339, 2003.

[18] A.G. Murzin, S.F. Brenner, T. Hubbard, and C. Chothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, 297:536–540, 1995.

[19] C.A. Orengo, A.D. Michie, S. Jones, D.T. Jones, M.B. Swindells, and J.M. Thornton. Cath- a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1108, 1997.

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

[21] A.A. Shelenkov, A.A. Shelenkov, and R.L. Dunbrak Jr. A graph-theory algorithm for rapid protein side-chain prediction. *Protein Science*, 9:2001–2014, 2003.

[22] M. J. Sippl. Calculation of conformational ensembles from potentials of mean force. *J. Mol. Biol.*, 213:859–883, 1990.

[23] M.J. Sternberg, P.A. Bates, K. A. Kelley, and R. M. MacCallum. Progress in protein structure prediction: Assessment of CASP3. *Curr. Opin. Struct. Biol.*, 9:368–373, 1999.