Predicting The β-Helix Fold From Protein Sequence Data

[Extended Abstract]

Phil Bradley* Lenore Cowen† ‡ Matthew Menke*

Jonathan King† § Bonnie Berger† *

ABSTRACT

A method is presented that uses β-strand interactions to predict the right-handed β-helix super-secondary structural motif in protein sequences. A program called `BetaWrap` implements this method, and is shown to score known β-helices above non-β-helices in the Protein Data Bank in cross-validation. It is demonstrated that `BetaWrap` learns each of the seven known SCOP β-helix families, when trained on the the known β-helices from outside the family. `BetaWrap` also predicts many bacterial proteins of unknown structure that play a role in human infectious disease to be β-helices; in particular, these proteins serve as virulence factors, adhesins and toxins in bacterial pathogenesis, and include cell surface proteins from Chlamydia and the intestinal bacterium *Helicobacter pylori*. The computational method used here may generalize to other β structures for which strand topology and profiles of residue accessibility are well conserved.

1. INTRODUCTION

Sophisticated algorithms (e.g., BLAST [1], FASTA [27]) exist for the detection of sequence similarity between proteins, and these provide the simplest and most commonly used tools for making structural and functional inferences about uncharacterized proteins. While impressive gains in sensitivity are being made by incorporating multiple aligned sequences (e.g., PSI-BLAST [2], HMMER [8], SAM-T98 [22]) and threading methods [31, 21, 7, 23, 33] have improved the detection of more distant homologies, there are still many examples of homologous proteins with strong structural similarities which cannot currently be detected. Here we introduce the `BetaWrap` program, which uses β-strand interactions to detect members of the parallel right-handed β-helix superfamily, a group of proteins characterized by widely divergent sequences but strong core structural similarities. The method may extend to other β-structures in which the strand topology and profiles of residue accessibility are well conserved.

It has been known for some time that in β-structural motifs amino acid residues that are close in space in the folded protein can exhibit marked statistical preferences [25, 17, 36]. These preferences have proven difficult to exploit, however, because residues in stacking β-strands that are close in 3D and may be instrumental in the fold, can be very far away in the 1D sequence. Thus in the absence of a related solved 3D structure or strong sequence homology with a solved 3D structure, it seems quite difficult to find the important correlations that could drive the fold. Of course, statistical correlations have long been used to predict secondary structure [12, 28], and the recent CASP3 competition [24] showed the strength of methods such as the PHD [28] and PSI-Pred [20, 19] programs in correctly locating α-helices and β-strands with over 70% accuracy from only the protein sequence.

Pairwise statistical correlations have been successfully employed to recognize α-helical super-secondary structural motifs, such as the two- and multi-stranded coiled-coil [6, 34, 29, 5, 30]. Aiding in the recognition of α-helical motifs, is the fact that residues that are close in space are also typically close in the one-dimensional sequence. For example, the methods of Berger et al. [6, 34, 29, 5, 30] were able to exploit the pairwise statistical correlations between residues in a short sliding window. As remarked above, it was less clear how to adapt these methods to any of the mainly-β super-secondary structural motifs. Even the ordinary secondary structure prediction methods are better at correctly placing α-helices, than β-strands [28].

The β-helix fold has a topology which makes prediction of the interacting residues in β-sheets more tractable. The fold is characterized by a repeating pattern of parallel β-strands in a triangular prism shape (Figure 1). The cross-section, or *rung,* of a β-helix consists of three β-strands connected by variable-length turn regions (Figure 2); the backbone folds

*Department of Mathematics and Lab for Computer Science, MIT, Cambridge, MA 02139, bab@mit.edu

†Corresponding Author

‡Department of Mathematical Sciences, Johns Hopkins University, Baltimore, MD 21218, cowen@cs.jhu.edu

§Department of Biology, MIT, Cambridge, MA 02139, jaking@mit.edu

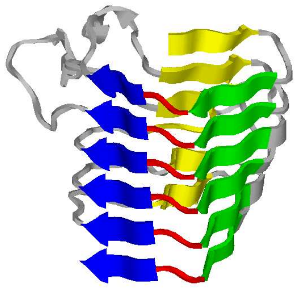

Figure 1: Side view of X-ray crystal structure of Pectate lyase C from Erwinia chrysanthemi [35], residues 102-258. β-sheet B1 is shown in light gray, B2 in medium gray, and B3 in black.

up in a helical fashion with β-strands from adjacent rungs stacking on top of each other in a parallel orientation. While the known β-helices vary in the number of complete rungs and in the lengths of the turn regions, the β-strand portions of the rungs have patterns of pleating and hydrogen bonding which are well-conserved across the superfamily [18].

Previous attempts at predicting β-helices have met with some success, but have not been successful at predicting β-helices across different families in the Protein Data Bank (PDB). Heffron et al. [14] developed a sequence-based profile from a pectate-lyase template, which failed to match any β-helices in the PDB other than the pectin and pectate lyases. Hidden Markov models for two β-helix families are in the PFAM database [8], which identify sequences only in those families. When we apply general sequence-based or threading methods to β-helices they primarily find with reasonable confidence levels sequences from the same family as the query sequence. A formal comparison with the iterative sequence-based method PSI-BLAST [1] and the publicly available threading program Threader [21] is described in the appendix.

Our algorithm `BetaWrap` is able to predict β-helices across all known families by generating and scoring different parses of the sequence into successive rungs. The scores computed are based on two main ingredients: a set of statistical preferences for aligned pairs in β-sheets, which was learned from a large database of amphipathic β-sheets; and a system of bonuses intended to reward for potential stacking interactions of the sort which are prevalent in the known β-helix structures.

`BetaWrap` scores the known β-helices ahead of all the non-β-helix proteins in a stringent cross-validation performed against a nonredundant version of the PDB. The β-helix superfamily is divided in the SCOP [26] database into seven families of closely related proteins.[1] Therefore, a sevenfold

[1]The structure of the Pectin Methylesterase protein from *Erwinia chrysanthemi* (PDB code 1QJV) was only recently solved and has not yet been placed in the SCOP database. Because of its low sequence and structural homology to the other known β-helices, we placed it by itself as one of the seven families.

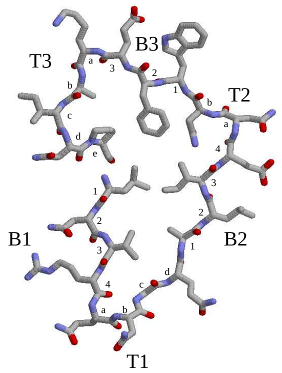

Figure 2: Top view of a single rung of a beta helix (residues 242–263) of Figure 1, parsed by the algorithm into β-strands B1, B2, B3 and the intervening turns T1, T2, and T3. Residues parsed as β-strand are numbered, and as turns are lettered. The alternating pattern of the strands before and after T2 is conserved across the superfamily.

cross-validation was performed, where all the proteins in the same family were left out of the training set in each experiment. Thus, `BetaWrap` was able to identify known β-helix proteins from one family, when only trained on β-helix proteins from a different SCOP family. In addition, the program makes reasonably good predictions of the alignment between sequence and structure in the known structures.

The `BetaWrap` program identifies a large set of sequences as having strong β-helical potential when run on the databases SWISS-PROT and TrEMBL ([3], see Section 4). As would be expected, some of the hits are proteins which are homologous or functionally equivalent to the known examples. A large number of pectate and pectin lyases are found, as well as additional polysaccharidases; the family of galacturonases is also well represented.

Perhaps most exciting, however, is the presence of a significant number of proteins which may have roles in pathogenesis; an example among the known structures is the P.69 pertactin from *Bordetella pertussis,* the toxin that causes Whooping Cough. Among the highest scoring of the protein sequences of unknown structure are several surface proteins from different *Chlamydia* species, and the intestinal bacterium *Helicobacter pylori.* Based on `BetaWrap` scores, we also predict a number of pollen allergens to be β-helices such as the Ragweed allergen. For a list of further interesting proteins of unknown structure that `Beta Wrap` predicts are β-helices, see Table 2. One of the most striking overall features of the set of proteins found is the non-random distribution of source organisms; from the roughly 600 proteins found in the TrEMBL database scoring above a high significance threshold, only a handful (13) of human and mouse proteins are found; none of these occurs within the 100 top-scoring proteins. Analysis of the identified sequences is underway.

2. THE ALGORITHM

The main component of the `BetaWrap` program is a novel "wrapping" algorithm that searches for the aligning parallel β-strands in successive rungs of the fold. While the turn lengths across different rungs of a β-helix can vary enormously (from a low of 2 residues to a high of 63 residues), the turn between β-strands B2 and B3 (the T2 turn, Figure 2) is more conserved; a majority of the rungs have a two residue turn at this location (with no known β-helix having fewer than six such rungs consecutively). More importantly, the hydrogen bonding and β-pleating patterns are conserved across these turns. Thus, given the sequence positions of two consecutive T2 turns in any of the known structures, one can say which residues are aligned and how they are oriented (relative to the core) in the strands which precede and follow the turns. [2] Consequently, the algorithm seeks to wrap a sequence of consecutive rungs with the T2 turn conserved; it locates the highest scoring wraps for a given amino acid sequence, as described below. The residues in the T2 turn are identified based on stacking preferences both in the turn, and in the surrounding residues from β-strands B2 and B3. The location of strand B1 is filled in to complete the parse of a generated wrap. Once the wraps are generated an α-helical secondary structure detector, based on an adaptation of the well-established GOR program [12], is applied as a filter to remove those which overlap with regions of high α-helix content.

2.1 First stage: the rungs subproblem

As a step toward the development of the wrapping algorithm, we first solve the following subproblem. Suppose we are given the amino acid sequence of a β-helix and told the sequence position of the T2 turn in one rung. Can we predict the location of the T2 turn in the next rung, assuming that both have exactly two residues?

The position of the second turn determines the residues that are in alignment in the two rungs. To score these aligned residue pairs, a database, called the *β-structure database* (see Section 3), of β-sheets was constructed which share with the β-helices the property that one face is buried and one exposed (the β-helices themselves were excluded from this database to avoid overtraining). The conditional probability that a residue of type X will align with residue Y, given their orientation relative to the core, was calculated from the β-structure database using standard methods (see, e.g. [4]). The natural logarithm of this probability gives the *pair score* of a vertical alignment of two residues. For a pair of aligned rungs, the *β-sheet alignment score* is the weighted sum of the seven alignment scores for the aligned pairs in the β-sheets B2 and B3 (a weight of 1 is given to the scores for inward pairs and 1/2 for the scores of the outward pairs, to reflect the fact that the environment of the inward residues is better conserved between β-helices than that of the outer pairs).

[2] We assume that the 3 residues following the turn and the 4 residues preceding it are participating in β-sheet interactions. While there are rungs in which this is not the case, the success of the algorithm indicates that these exceptions do not pose a significant problem.

The β-sheet alignment score is the heart of the recognition method; however, we improve its performance with several bonuses and penalties:

- Based on the β-helix structures in the training set, a distribution on turn lengths was learned. The first adjustment to the score is a gap penalty that penalizes alignments that leave too many or too few residues unmatched between two rungs (based on standard deviations from the mean).
- A bonus is added when two aromatic amino acid residues (F, Y, or W) appear stacked on top of each other, when a stack of β-branched aliphatic residues (V or I) are seen in alignment, and for the inward pointing polar residues (C,S,T,and N) seen to stabilize the T2 turn by forming hydrogen bonds in the known structures. These stacking preferences are based on the interactions prevalent in the known structures; for a full discussion see [18].
- A strong penalty effectively disallows the highly charged residues (D,E,R, and K) from appearing in the inward-pointing positions of a β-strand (see [14]).

The success of this scoring system at identifying rung-rung pairs is described in Section 4.

2.2 From a rung to multiple rungs

To adapt the rung-to-rung scoring system of the previous section to the problem of generating complete wraps, initial B2-T2-B3 segments must be located. Here a simple sequence template is used, based on the assumption that hydrophobic residues (plus tyrosine, which is often found in the interior of the β-helices) will appear at the inward positions of the β-sheets. Thus the initial rungs are simply matches to the pattern: $\Phi X \Phi X X \phi X \Phi$, where Φ matches one of the residues (A,F,I,L,M,V,W, or Y), ϕ matches any amino acid except (D,E,R, or K), and X matches any amino acid at all.

Beginning with each substring that matches this pattern, the five top scoring aligned rungs are calculated both forward and backward in the sequence. This process is repeated with each of these rungs, and with their aligned rungs, continuing until a tree of potential 5-rung wraps extending both forward and backward in sequence is generated. In this way, the B2-T2-B3 portions of wraps containing each of the initial rungs are generated; this phase of the algorithm is optimized using dynamic programming. The score attached to a given wrap is the average of its rung-to-rung alignment scores. The collection of wraps is subject to three stages of filtering, as described in the next two sections (the cutoffs for these filters are recalculated in each of the cross-validation runs based on the training data for that run; see Section 3). The *wrap score* assigned to a candidate β-helix is the average of the scores for the top ten wraps which pass this filtering stage. Averaging the top ten wrap scores rules out spurious hits to sequences in which a single high-scoring wrap is found by chance (when applied to the known β-helices, the algorithm produces a large number of high-scoring wraps: the correct wraps, but also many mostly-correct wraps with comparable scores). If less than ten wraps remain after filtering the protein is rejected.

2.3 Completing the parse

Although the relative positioning of the rungs in a wrap is fixed by the above procedure, the positions of the B1 strands are not determined. The algorithm scores potential placements of the B1 strands into the parse using the same strand-strand alignment scores described above (β-alignment probabilities and stacking bonuses); the process is guided by a second gap score learned from the distributions of the T1 turn lengths in the known structures (there is a marked preference for T1 turns of length three, four, and five). The highest scoring B1 parse is chosen for the wrap. Note that the score for this B1 parse does not change the score of the wrap; however, a wrap is rejected if a B1 parse scoring above a predetermined threshold cannot be found.

Once the complete wraps are generated, they are filtered based on residues found at two positions in the turns. The a positions of the T1 and T2 turns (Figure 2) show distinctive residue preferences, in particular the larger hydrophobics (V,I,L,F,M,and W) are strongly disfavored, and these preferences run counter to what would be expected if the pleating pattern of the preceding strands is extended forward. As described more fully in Jenkins et al. [18], the a position of T2 has unique structural features (most notably an α_L conformation) which constrain the types of residues found there (no large hydrophobics are found at this position in any of the rungs of the known structures). As a consequence, a wrap is only permitted a single large hydrophobic at the T2 a position; in addition, if the total number of hydrophobics at both a positions in T1 and T2 exceeds a predetermined cutoff, a penalty is assessed. This has the effect of penalizing spurious matches to proteins which have longer β-strands than those found in the β-helices.

2.4 The α-helical filter

The information-theoretic methodology of GOR-IV [12] was adapted to construct a two-state (α-with-high-confidence/other) secondary-structure predictor (details in full version of the paper). GOR-IV was used in preference to more recent algorithms, e.g. those using multiple sequence information, because its simple statistical framework and single-sequence input was easy to specialize for our purpose: the prediction of regions of high α-content. Wraps were filtered on the basis of their predicted α-content, with the aim of removing β-helix parses which overlap with all-α regions. If the total fraction of a wrap predicted as α-helical with high confidence exceeds a threshold the wrap is rejected. In addition each of the four complete internal rungs are judged on the basis of their fraction of predicted α-content, and if more than two are rejected, the wrap as a whole is rejected. Here a rung is rejected if it contains at least four predicted α-residues and has a fraction of α-content exceeding a second threshold. As described in Section 3, these thresholds are based on the training data and thus were recalculated for each of the cross-validation runs.

Sequences are prefiltered for trans-membrane α-helices using the GES hydrophobicity scale [10], a window of size 21, and a threshold of -2 kcal/mol. The predicted helices are removed, and the query sequence is broken into subsequences which are scored individually.

3. METHODS

3.1 The databases

The PDB-minus database was constructed from the PDB_select 25% list of June 2000 [16, 15], with the β-helices removed. (PDB_select is a subset of the PDB in which no two proteins have sequence similarity greater than a cutoff, in this case, 25%.) The database contained 1346 sequences.

The β-structure database was constructed from PDB-minus (with membrane proteins removed) by looking for alternating patterns of residue accessibility in β-strands. The PDB-minus structure files were processed using the program Stride [11], which annotates secondary structure, hydrogen bonds, and residue accessibilities. β-sheets whose residue accessibilities fit an alternating pattern of buried/exposed were identified, and the aligned residue pairs were annotated based on the hydrogen bonding patterns. In all, 650 protein chains from PDB-minus contributed sheets or portions of sheets to the database.

New β-helices were identified from the sequence databases SWISS-PROT (Release 39.6 of 30-Aug-2000: 88166 entries) and TrEMBL (Release 14.11 of 25-Aug-2000: 301497 entries) [3].

3.2 Training

A seven-fold cross-validation was performed on the seven β-helix families of closely related proteins in the SCOP [26] database. (The structure of the Pectin Methylesterase protein from *Erwinia chrysanthemi*, PDB code 1QJV, was only recently solved and has not yet been placed in the SCOP database. Because of its low sequence and structural homology to the other known β-helices, we placed it by itself as one of the seven families.) PDB-minus was randomly partitioned into a 60% training (with 815 structures) and 40% (with 531 structures) testing set. For each cross, proteins in one β-helix family were placed in the test set, while the remainder of the beta helices were placed in the training set. The scores reported for the β-helix proteins in Table 1 and in Figure 3 are the scores in the leave-family-out cross experiment for that β-helix's protein family. The optimal thresholds for the α-filter, the distribution of the gap penalties (as described in Section 2), the B1-score threshold, and the hydrophobic-count threshold were optimized for training data, and thus recalculated for each experiment.

4. RESULTS

There is no overlap in the scores computed by `BetaWrap` when the histogram scores for the β-helix database are plotted against those for the PDB-minus database (Figure 3). The score for each β-helix is taken from its cross-validation run. In Table 1, the β-helix proteins used in this study are listed along with their cross-validation scores and ranks, as compared with the other members of their SCOP family and the sequences in PDB-minus. The three top-scoring non-β-helix proteins are the coat protein (4SBV:A) from Southern Bean Mosaic Virus (an eight-stranded β-sandwich) with a score of -20.78; Tetrahydrodipicolinate N-Succinyltransferase (3TDT) from Mycobacterium bovis (a left-handed parallel β-helix) with a score of -20.83; and Vp1 protein (1B35:C) from Cricket Paralysis Virus (another eight-stranded

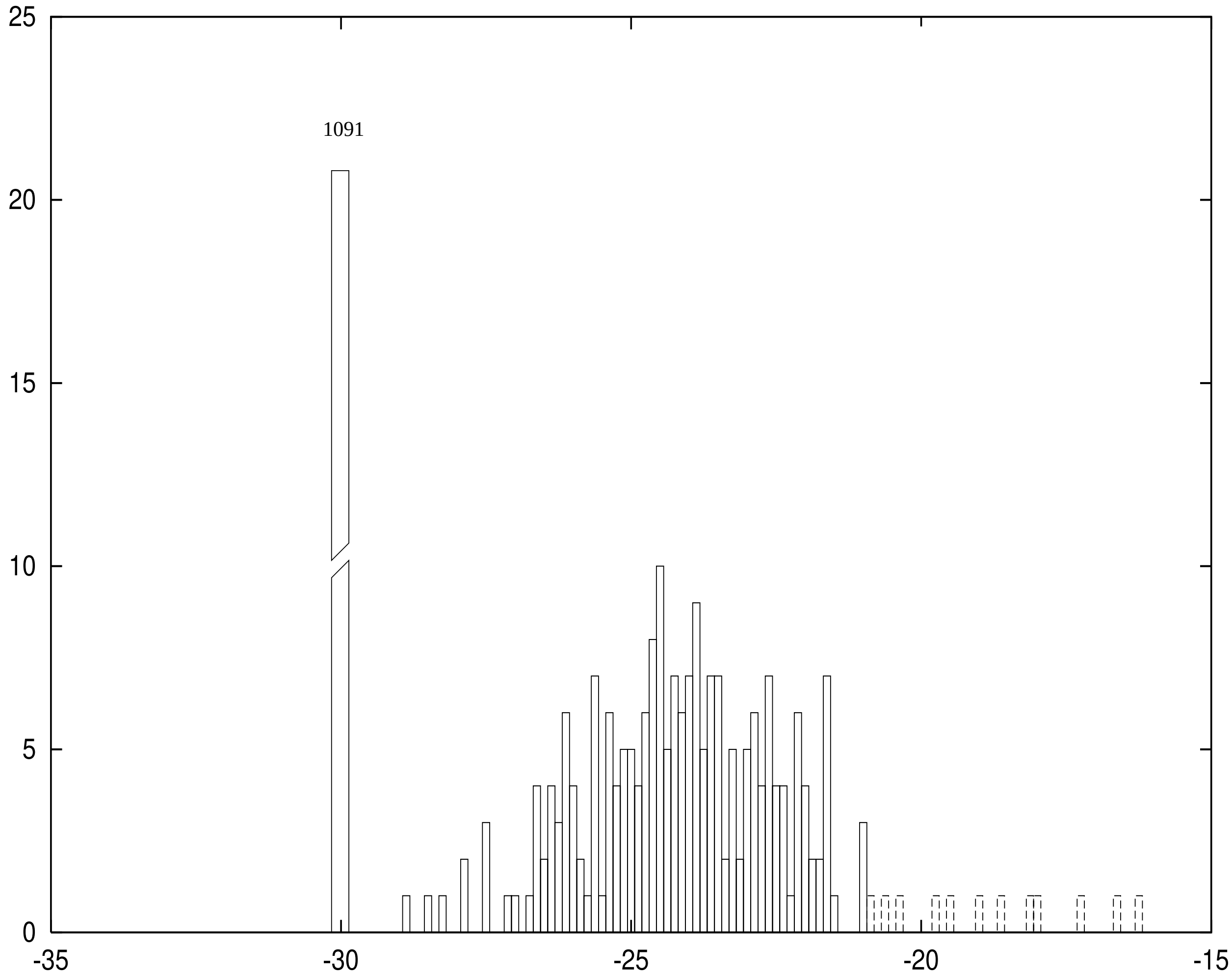

Figure 3: Histogram of protein scores as computed by `BetaWrap`. The β-helix scores (12 proteins) were superimposed on the scores of the PDB-minus database (1346 proteins), with the 1091 proteins which could not be successfully wrapped (Section 2.2) given the arbitrary score -30. The β-helix histogram is dashed, and PBD-minus is solid.

SCOP Family	Name	Source	PDB	Rank	Score
Pectate Lyase	Pectate Lyase E	*Erwinia chrysanthemi*	1PCL	1	-16.02
Pectate Lyase	Pectate Lyase C	*Erwinia chrysanthemi*	1PLU	2	-16.44
Pectate Lyase	Pectate Lyase	*Bacillus subtilis*	1BN8	3	-18.42
Pectin Lyase	Pectin Lyase B	*Aspergillus niger*	1QCX	1	-17.09
Pectin Lyase	Pectin Lyase A	*Aspergillus niger*	1IDK	2	-17.99
Galacturonase	Polygalacturonase	*Erwinia carotovora*	1BHE	1	-18.80
Galacturonase	Polygalacturonase II	*Aspergillus niger*	1CZF	2	-19.32
Galacturonase	Rhamnogalacturonase A	*Aspergillus aculeatus*	1RMG	3	-20.12
P22 Tailspike	P22 Tailspike	*S. typhimurium* Phage P22	1TSP	1	-20.46
P.69 Pertactin	P.69 Pertactin	*Bordetella pertussis*	1DAB	1	-17.84
Chondroitinase	Chondroitinase B	*Flavobacterium heparinium*	1DBO	1	-19.55
Unclassified	Pectin Methylesterase	*Erwinia chrysanthemi*	1QJV	1	-20.74

Table 1: Known β-helices and their `BetaWrap` scores/ranks

β-sandwich from the same SCOP superfamily, viral coat and capsid proteins, as 4SBV:A).

4.1 Predicted alignments between sequence and structure

As well as its strong success in predicting the presence or absence of the β-helix motif, the algorithm shows some success in predicting the location of the rungs in the known β-helices. Nine of the twelve proteins have a correct wrap of the B2-T2-B3 region within the 10 scored parses. The other three proteins, 1TSP, 1CZF, and 1QJV, have wraps with the correct placement of two, three, and four of the five rungs, respectively. The protein wrapped with greatest success is the galacturonase 1BHE, with four of the 10 scored parses correct and the other six off in a single rung.

This success at generating complete wraps follows from the accuracy of the program in predicting rung-rung alignments. As described in Section 2.1, a rung-rung alignment score was developed to predict the location of the next rung in a wrap given the previous one. When tested on the 77 rung-rung pairs in the known structures for which both rungs have a two residue T2, the correct alignment of the second rung with the first is given the highest score in 58 pairs. Furthermore, the correct alignment appears in the top five scoring alignments in 72 of the 77 pairs (recall that the five top scoring aligned rungs are kept at each stage in generating the tree of wraps from an initial rung).

4.2 New β-helix candidates

The `BetaWrap` program has identified many new sequences that we believe contain β-helix structures. Table 2 lists some examples of the predicted proteins. A number of these are functionally similar to the known β-helices. The protein from *R. leguminosarum* is a polysaccharidase, and the bacteriophage tail protein has features in common with the P22 tailspike (R. Seckler, personal communication). Two of the proteins, WCAM from *Salmonella typhimurium* and the hypothetical product of the SPSR gene in *Sphingomonas sp. S88*, are involved in polysaccharide synthesis, and several are surface proteins which may have roles in virulence. The *B. pertussis* protein BRKA was also predicted to have a β-helical structure by Emsley et al. [9] based on sequence similarity to P.69 pertactin. For a more complete list, see `http://theory.lcs.mit.edu/betawrap`. There is a definite bias in the distribution of source organisms among the high-scoring proteins. Very few human, fly, or mouse proteins are found in spite of their over-representation in the databases. This is in agreement with the observed species distribution of the known β-helices.

5. DISCUSSION

Our results indicate that there are correlations in β-structures and features of β-helices that can help distinguish the parallel right-handed β-helix from non-β-helix domains. It is possible that there are structural features of the β-helices in our database that are not general features of β-helices. Even within the known structures, however, there is sufficient variation to suggest the robustness of the algorithm; for example, the program successfully wraps even those β-helices (such as 1RMG, see Table 1) which have an additional β-strand inserted between B1 and B2. In addition, the relative success of our β-helix prediction method in identifying plausible new candidates for β-helices suggests that inherent biases are not great.

While the program does achieve complete separation of the β-helix scores from those of PDB-minus, it is likely that there will be non-β-helices in larger sequence databases whose scores under the current algorithm overlap with those of the lowest scoring β-helices. There are a number of directions being explored to improve the confidence of predictions in this score range. One possibility is to incorporate evolutionary information about a query sequence in the scoring procedure (significant gains have been made when such information is used in secondary-structure prediction). The algorithm could take as input a multiple alignment of homologous sequences, scoring whole columns rather than the individual residues of a query sequence. An alternative (which would not be as sensitive to the accuracy of the alignments) would be to score single sequences but then consider the ensemble of scores for all proteins (or domains) within a family (such as those collected in Pfam [32]). These methods would likely aid in finding new families of β-helices for which the scores of the individual members are borderline, and in eliminating single proteins which score highly by chance, as the features which produce the score are unlikely to be conserved in homologs. Another possibility is the use of 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]).

Work is also under way to improve the sequence-to-structure alignments produced by the algorithm. A second stage is being implemented to extend the predicted wraps (which in all cases represent only a portion of the helical structure) outward to give complete folds. It will probably be necessary to relax the turn-length restrictions in order to guarantee that we can find these additional rungs. Correct alignments of the newly discovered β-helices will hopefully be useful in predicting functional residues, and in designing mutational studies which could in turn lend support to the prediction. A large class of mutations that affect the folding and stability of the P22 Tailspike protein have been identified and characterized, and there is evidence that the β-helix domain is particularly sensitive to mutations affecting its folding ([13]).

We hope that the methods described here can be applied to other families of β-structure. It is plausible that one could achieve similar results by modifying the wrapping algorithm to reflect a different strand topology and gap distribution, and replacing the bonuses particular to β-helices with a set learned from the new set of structures.

6. BETAWRAP ON THE WEB

A server running `BetaWrap` is available on the Internet, at `http://theory.lcs.mit.edu/betawrap`. This site also contains an updated list of high-scoring protein sequences.

7. ACKNOWLEDGMENTS

This work was supported in part by the Charles E. Reed Faculty Initiatives Fund. B.B. was supported in part by an NSF Career Award. J.K. was supported in part by NIH GM 17980. P.B. was supported by an MIT/Merck graduate fel-

ID	Description	Organism	Score
P74816	Hypothetical 69.5 KDA protein Gene: SPSR	*Sphingomonas sp.* S88	-13.85
O64135	YORA protein	Bacteriophage SPBc2	-14.05
O05692	Polysaccharidase	*Rhizobium leguminosarum*	-14.64
O25579	Toxin-like outer membrane protein	*Helicobacter pylori*	-16.05
190K_RICRI	190 KDA antigen precursor	*Rickettsia rickettsii*	-16.86
OMPF_CHLTR	Putative outer membrane protein F precursor	*Chlamydia trachomatis*	-17.08
CSG_METSC	Cell surface glycoprotein precursor (S-layer protein)	*Methanothermus sociabilis*	-17.90
MPA2_AMBAR	Pollen allergen AMB A 2	*Ambrosia artemisiifolia*	-18.11
Q9ZGR4	Putative cytotoxin (Gene L7095)	*Escherichia coli* O157:H7	-18.69
WCAM_SALTY	Colanic acid biosynthesis protein WCAM	*Salmonella typhimurium*	-19.04
TSPE_BPSFV	Bifunctional tail protein	Bacteriophage SfVI	-19.31
Q45340	BRKA	*Bordetella pertussis*	-20.21

Table 2: Examples of proteins predicted to form β-helices by `BetaWrap` with their scores. Identifiers (ID) and descriptions are taken from SWISS-PROT or TrEMBL. For an updated list of proteins with high `BetaWrap` scores, see `http://theory.lcs.mit.edu/betawrap`
.

lowship. L.C. was a visitor at MIT supported by a Emmaline Bigelow Conland Fellowship at the Radcliffe Institute for Advanced Study. Thanks to Scott Decatur, Jonathan Dunagan, and Brian Dean for their initial computational assistance and to Russell Schwartz, Peter Thumfort, Sara Little, Jesse Barnes, and Patricia Clark in the King lab for many helpful comments that greatly improved the quality of this work.

Appendix: Comparison with other methods

Two existing methods were examined for their ability to detect the relationships between the known families of β-helices. First, the sequences of the 12 β-helix domains were used to search the NCBI nonredundant database (14-Dec-2000 update, 595890 entries) using the iterative multiple sequence alignment program PSI-BLAST [1] (version 2.1.2). The default e-value threshold for inclusion of 0.001 was used; all searches converged before 20 rounds. A sequence was considered as having been found if it was included in the profile after any of the rounds.

Four of the sequences gave profiles which included only a single sequence of known structure, the initial query sequence; these sequences were not found in searches with other β-helix sequences. When sequences from the remaining three families were used as queries, cross-family relationships were detected. In particular, pectate lyases were found from pectin lyase queries, and visa versa, and each of the galacturonase sequences found either some of the pectate or some of the pectin lyase sequences as well (Table 3).

Next, the program Threader 2.5 [21] was used to thread the 12 β-helix sequences onto an accompanying fold library (3-99 version, 1906 domains). Threadings were sorted by the Z-scores of the combined pairwise and solvation energies and filtered using the core-shuffled pairwise energies, as described in the user manual. The most recent available fold library contains three β-helix structures: 1PLU, 1RMG, and 1TSP. The five pectate and pectin lyase sequences were matched to the 1PLU template with highest confidence. Matches to the other two templates scored lower than threadings onto non-β-helices. 1RMG and 1CZF were matched to the 1RMG template with highest confidence; again matches to the other templates scored lower than threadings onto non-β-helices. 1TSP was threaded onto its structure with highest confidence but did not match the other two templates. Matches of the remaining sequences with the three templates all scored lower than threadings onto non-β-helices. Thus Threader was able to recognize the similarity of the pectin lyases to the pectate lyase 1PLU, but did not recognize other cross-family similarities. One can reasonably conclude that there are additional β-helices in the sequence databases which would not be detected by either of these methods.

REFERENCES

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

	1PLU	1PCL	1BN8	1IDK	1QCX	1RMG	1BHE	1CZF	1TSP	1DAB	1DBO	1QJV
1PLU	X	X	X	X	X							
1PCL	X	X	X	X	X							
1BN8	X	X	X	X	X							
1IDK	X	X	X	X	X							
1QCX	X	X	X	X	X							
1RMG	X	X			X	X	X	X				
1BHE	X	X	X	X	X	X	X	X				
1CZF					X	X	X	X				
1TSP									X			
1DAB										X		
1DBO											X	
1QJV												X

Table 3: Results of PSI-BLAST searches on the known β-helix structures. An 'X' indicates that the protein in that column was found when searching with the protein indexing the given row. SCOP families are separated by horizontal lines. While PSI-BLAST finds pectate lyases from pectin lyases, and vice versa, and also finds pectate and pectin lyases sequences from some of the galacturonases (but not vice versa) the remaining four sequences were not matched to or by other β-helices in the searches described above.

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

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

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

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

[6] B. Berger, D. B. Wilson, E. Wolf, T. Tonchev, M. Milla, and P. S. Kim. Predicting coiled coils by use of pairwise residue correlations. *Proc. of the Natl. Academy of Sci., USA*, 92:8259–8263, Aug. 1995.

[7] S. Bryant. Evaluation of threading specificity and accuracy. *Proteins*, 26:172–185, 1996.

[8] S. Eddy. Hidden Markov models and large-scale genome analysis. *Transactions of the American American Crystallographic Association*, 1997.

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

[10] D. Engelman, T. Steitz, and A. Goldman. Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins [review]. *Annual Review of Biophysics and Biophysical Chemistry*, 15:321–53, 1986.

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

[12] J. Garnier, J. Gibrat, and B.Robson. GOR secondary structure prediction method version IV. *Methods in Enzymology*, 266:540–553, 1996.

[13] C. Haase-Pettingell and J. King. Prevalence of temperature sensitive folding mutations in the parallel beta coil domain of the phage p22 tailspike endorhamnosidase. *J. Mol. Biol.*, 267:88–102, 1997.

[14] S. Heffron, G. Moe, V. Sieber, J. Mengaud, P. Cossart, J. Vitali, and F. Jurnak. Sequence profile of the parallel β helix in the Pectate Lyase superfamily. *Journal of Structural Biology*, 122:223–235, 1998.

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

[16] U. Hobohm, M. Scharf, R. Schneider, and C. Sander. Selection of a representative set of structures from the Brookhaven Protein Data bank. *Protein Science*, 1:409–417, 1992.

[17] T. Hubbard and J. Park. Fold recognition and ab initio structure predictions using hidden Markov models and beta-strand pair potentials. *Proteins*, 3:398–402, 1995.

[18] J. Jenkins, O. Mayans, and R. Pickersgill. Structure and evolution of parallel β-helix proteins. *Journal of Structural Biology*, 122:236–246, 1998.

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

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

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

[22] K. Karplus, C. Barrett, and R. Hughey. Hidden Markov models for detectin remote protein homologies. *Bioinformatics*, 14:846–856, 1998.

[23] L. Kelley, R. MacCallum, and M. Sternberg. Enhanced genome annotation using structure profiles in the program 3D-PSSM. *J. Mol. Biol.*, 299(2):501–522, 2000.

[24] P. Koehl and M. Levitt. A brighter future for protein structure prediction. *Nat. Struct. Biol.*, 6:108–111, 1999.

[25] S. Lifson and C. Sander. Specific recognition in the tertiary structure of β-sheets of proteins. *Journal of Molecular Biology*, 139:627–629, 1980.

[26] A. Murzin, S. 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.

[27] W. Pearson and D. Lipman. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA*, 85:2444–8, 1988.

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

[29] M. Singh, B. Berger, and P. S. Kim. Learncoil–VMF: Computational evidence for coiled coil-like motifs in many viral membrane fussion proteins. *J. of Molecular Biology*, 290(1):241–251, 1999.

[30] M. Singh, B. Berger, P. S. Kim, J. Berger, and A. Cochran. Computational learning reveals coiled coil-like motifs in histidine kinase linker domains. *Proc. of the Natl. Academy of Sci., USA*, 95(6):2738–2743, 1998.

[31] M. Sippl and S. Weitckus. Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a data base of known protein conformations. *Proteins*, 13:258–271, 1992.

[32] E. Sonnhamer, S. Eddy, E. Birney, A. Bateman, and R. Durbin. Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res*, 26(1):320–322, 1998.

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

[34] E. Wolf, P. S. Kim, and B. Berger. `MultiCoil:` a program for predicting two and three stranded coiled coils. *Protein Science*, 6(6):1179–1189, June 1997.

[35] M. D. Yoder, N. T. Keen, and F. Jurnak. New domain motif: structure of pectate lyase C, a secreted plant virulence factor. *Science*, 260:1503–1507, 1993.

[36] H. Zhu and W. Braun. Sequence specificity, statistical potentials and 3D structure prediction with self-correcting distance geometry calculations of beta-sheet formation in proteins. *Protein Science*, 8:326–342, 1999.