JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 9, Number 2, 2002
© Mary Ann Liebert, Inc.
Pp. 261–276

Predicting the Beta-Helix Fold from Protein Sequence Data

LENORE COWEN,[1] PHIL BRADLEY,[2] MATTHEW MENKE,[2] JONATHAN KING,[3]
and BONNIE BERGER[3]

ABSTRACT

A method is presented that uses β-strand interactions to predict the parallel 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 primarily on β-structures that are not β-helices, together with structural features of known β-helices from outside the family. BetaWrap also predicts many bacterial proteins of unknown structure 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.

Key words: parallel right-handed beta helix, protein folding, protein motif recognition, statistical prediction, bacterial pathogenesis.

1. INTRODUCTION

This paper presents the first computational method that successfully predicts membership in a β-structural super-secondary fold family based on a protein's amino acid sequence. We introduce the BetaWrap program, which recognizes sequences whose folds belong to the parallel right-handed β-helix superfamily. There currently exist 12 known examples of the right-handed β-helix whose crystal structures have been solved in the Protein Data Bank (PDB) (Berman *et al.*, 2000). The SCOP (Murzin *et al.*, 1995) classification system (version 1.53) places these known β-helices into seven different families and though there are core structural similarities that categorize the β-helix, there is little sequence homology across structures in different families (except the pectate and pectin lyases). Thus even multiple-alignment sequence-based methods such as PSI-BLAST (Altschul *et al.*, 1997) cannot solve the β-helix recognition problem across many of the different SCOP families. Threading methods (Sippl and Weitckus, 1992; Jones *et al.*, 1992; Bryant, 1996; Kelley *et al.*, 2000; Sternberg *et al.*, 1999) and hidden Markov methods such

[1] Department of EECS, Tufts University, Medford, MA 02155.

[2] Department of Mathematics and Laboratory for Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139.

[3] Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139.

as HMMER (Eddy, 1998) also do not solve the β-helix recognition problem across the different SCOP families; see Section 5.

It has been known for some time that amino acid residues that are close in space in a folded three-dimensional protein structure can exhibit marked statistical preferences. For example, consider the coiled-coil α-helical protein motifs. The backbone coils in a processive turn, giving a structural repeat every seven residues. A residue is closest in space to the residues that appear +1, +3, and +4 away in the sequence, and strong pairwise statistical correlations between residues appearing in these positions in the coiled coil were identified by Berger *et al.* (1995). Because residues a fixed distance in sequence exhibit these correlations in coiled coils Berger *et al.* (Berger *et al.*, 1995; Wolf *et al.*, 1997; Singh *et al.*, 1999; Berger and Singh, 1997; Singh *et al.*, 1998), could predict the coiled-coil fold from the amino acid sequence alone, by sliding a 30-residue window along the sequence and tabulating pairwise frequencies in these positions, as compared to the pairwise frequencies in positive and negative examples of known coiled coils. Regions of α-helix secondary structure (with a slightly less regular repeat) have also been successfully predicted by these methods, again based on the fact that residues close in space are also close in sequence (Garnier *et al.*, 1996).

In β-structural motifs, as was noted by Lifson, Sanders and others (Lifson and Sanders, 1980; Hubbard and Park, 1995; Zhu and Braun, 1999), amino acid residues that are close in space also exhibit marked statistical preferences. 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 (and a variable number of amino acid residues apart) 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. Even the best (local) secondary structure predictors such as PHD (Rost and Sander, 1993) and PSI-Pred (Jones, 1999a,b) are better at correctly placing α-helices than β-strands (Rost and Sander, 1993).

We chose to target the β-helix as a good first β-structural motif to attempt to predict computationally because it has a topology that makes prediction of the interacting residues in the β-sheets more tractable, namely a long, processive fold. The fold is characterized by a repeating pattern of parallel β-strands in a triangular prism shape (Yoder *et al.*, 1993) (Fig. 1). The cross-section, or *rung,* of a β-helix consists of three β-strands connected by variable-length turn regions (Fig. 2); the backbone folds 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 that are well conserved across the superfamily (Jenkins *et al.*, 1998). Examination of the solved β-helix structures (Yoder and Jurnak, 1995; Kreisberg *et al.*, 2000) together with the analysis of mutants defective in the folding of β-helices (Kreisberg *et al.*, 2000) suggested that the interaction of the strand side-chains in the buried core

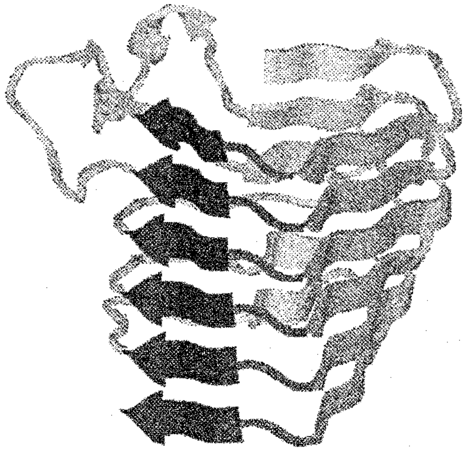

FIG. 1. Side view of X-ray crystal structure of Pectate lyase C from *Erwinia chrysanthemi* (Rost and Sander, 1994), residues 102-258; β-sheet B1 is shown in light gray, B2 in medium gray, and B3 in black.

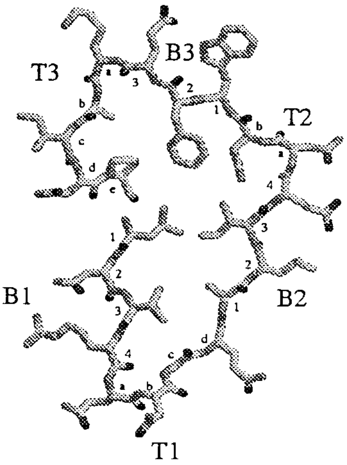

FIG. 2. Top view of a single rung of a β-helix (residues 242–263) of Fig. 1, parsed by the algorithm into β-strands B1, B2, B3 and the intervening turns T1, T2, and T3. Residues parsed as β-strand are numbered and as turns are lettered. The alternating pattern of the strands before and after T2 is conserved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

were critical determinants of the fold. We incorporated this interior packing emphasis in the development of the program.

The method we employ has two phases, a wrapping phase and a scoring phase. The wrapping phase can best be understood with an analogy with old historical ciphers. As documented as far back as Ancient Greece, secret coded messages were once sent as a strip of letters, that, when wrapped around a tube of the appropriate diameter, would align vertically to spell out words of the message. Wrapped incorrectly, they would spell out nonsense (see Fig. 3). In the wrapping phase, we do much the same thing with an amino acid sequence; however, the complexity is increased by the fact that there are loops of various lengths, so the sequence does not wrap tight. The algorithm tries to wrap the unknown sequence in all plausible ways and see if any of these wraps make "sense." The "sense" is determined by the scoring phase whose principal component is based, just as in the coiled-coil algorithm, on examining pairwise frequencies of specific pairs of amino acids. Instead of these pairs being a fixed distance apart in sequence, they now are the pairs that stack in the 3D wrapping (and may, in fact, be far away and a variable distance apart

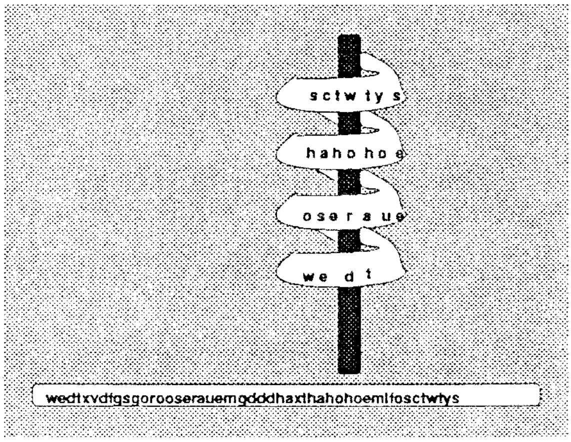

FIG. 3. An ancient cipher. The strip of letters when wrapped around the dowel can be made to spell out vertically the words "show case the word that you see." Our method wraps up a sequence (with a variable width dowel) in all possible ways to see if it can attain plausible vertical alignments.

in sequence.) Thus, our method can be seen as a spatial generalization of the PairCoil method of Berger *et al.* (1995).

Our BetaWrap program 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 (Murzin *et al.*, 1995) database into seven families of closely related proteins.[1] Therefore, a sevenfold cross-validation was performed, where all the proteins in the same family were left out of the training set in each experiment. 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.

We remark that even though BetaWrap has access to the other six β-helix families in each cross-validation, it makes very parsimonious use of their sequence information. The β-helices in the training set yield the structural template, and also set cutoffs for the gap lengths between β-strands, and a filter for α-helical content. The stacking preferences that are at the core of the method are initially built from a database of amphipathic β-structures that are not β-helices (see Section 3).

In comparison, we ran the iterative sequence-based method PSI-BLAST (Altschul *et al.*, 1990), the publicly available threading program Threader (Jones *et al.*, 1992), and the hidden Markov model program HMMer (Eddy, 1998), to see if they could predict any of the known β-helices from given known β-helix sequences. Both PSI-BLAST and Threader primarily find, with reasonable confidence levels, sequences from the same family as the query sequence. The only exception is that there is some cross-over between the pectin lyase and pectate lyase families (which have known sequence similarities) plus, when PSI-BLAST is given a query sequence from the galacturonase family, it picks up some of the pectin lyase and pectate lyase sequences (but not vice versa). Heffron *et al.* (1998) developed a sequence-based profile from a pectate lyase template to try to predict the β-helix specifically. Their template failed to match any β-helices in the PDB other than the pectin and pectate lyases; however, it predicted many pectate lyases to be β-helices in the larger databases of protein sequences of unknown structure. HMMer (Eddy, 1998), tested with structural multiple alignments of six β-helix families based on seven different FSSP seeds (Holm and Sander, 1996), one for each β-helix family, also picks up pectate and pectin lyase families, and one of the galacturonases (1RMG), but fails on the other SCOP families to separate β-helix from non-β-helix structures (see Section 5).

The BetaWrap program also identifies a large set of sequences as having strong β-helical potential when run on the databases SWISS-PROT and TrEMBL (Bairoch and Apweiler, 2000; see Section 4). It scores 2,248 proteins higher than the lowest-scoring β-helix when searching the 595,890 sequences in the NCBI non-redundant database. (Table 4 in Section 4.2 will list a few of the top-scoring proteins that we strongly predict to have a β-helical structure based on their performance using 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. More information can be found on our website at *theory.lcs.mit.edu/betawrap.*

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, Fig. 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

[1] The structure of the pectin methylesterase protein from *Erwinia chrysanthemi* (PDB code 1QJV) was not yet solved and had not yet been placed in the SCOP database (in version 1.53). 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, as do newer versions of SCOP.

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 (Garnier *et al.*, 1996), is applied as a filter to remove those which overlap with regions of high α-helix content. A transmembrane filter is also applied.

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, divided by the frequency of residues of type X, given their orientation relative to the core, was estimated from the β-structure database using standard methods (see, e.g., Berger [1995]). 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 β-structure database have been reproduced below in Tables 1 and 2. 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).

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

- 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). The gap penalty is given by $\lfloor|(\text{gap} - 19)/6|\rfloor$, where "gap" is the number of residues inserted between the B3 strand of the first rung in the sequence and the B2 strand of the second. (The parameters 19 and 6 are based on the gap distribution for the known β-helices.)
- For internal residues, a bonus of +2 is added when two aromatic amino acid residues (F, Y, or W) appear stacked on top of each other, +1 when a stack of β-branched aliphatic residues (V or I) are seen in alignment, and +1 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. Aligned asparagines at the inward T2 position receive an additional +1 bonus. These stacking preferences are based on the interactions prevalent in the known structures; for a full discussion see Jenkins *et al.* (1998).
- If a highly-charged residue from the set (D, E, R, and K) appears in the inward-pointing positions of a β-strand, the parse is discarded (see Heffron *et al.*, 1998).

When tested on the 77 rung–rung pairs in the known structures for which both rungs have a two residue T2 turn, the correct alignment of the second rung with the first is given the highest score in 58 pairs.

[2] We assume that the three residues following the turn and the four 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.

Table 1. Conditional Probabilities for Alignment of Buried Residues from the β-Structure Database[a]

	A	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*K*	*L*	*M*	*N*	*P*	*Q*	*R*	*S*	*T*	*V*	*W*	*Y*
A	5.8	5.5	8.9	19.5	7.5	3.6	12.6	7.8	5.5	8.8	5.4	10.0	6.9	4.6	2.9	6.7	8.4	7.8	7.9	7.4
C	2.2	11.1	3.5	2.4	2.8	4.3	3.1	2.5	5.5	2.8	1.4	3.3	13.9	2.3	2.9	4.3	4.2	2.2	6.1	3.0
D	1.0	1.0	0.8	0.8	0.6	2.1	1.5	0.8	0.8	0.3	0.9	5.0	4.6	0.8	5.8	3.0	0.6	0.8	0.8	0.3
E	1.6	0.5	0.6	4.8	0.9	0.3	3.1	0.6	11.1	0.4	1.4	0.6	0.6	0.6	0.6	1.2	0.6	0.1	0.6	0.6
F	9.9	9.0	7.1	14.6	11.9	13.8	11.1	8.4	11.1	9.1	11.8	5.0	11.6	13.9	8.8	14.1	11.4	9.0	8.8	11.1
G	2.0	6.0	10.7	2.4	5.9	6.5	7.9	3.6	4.3	3.6	2.9	1.6	13.9	2.3	2.9	3.0	6.0	4.4	2.6	5.2
H	1.6	1.0	1.7	4.8	1.1	1.8	6.3	0.6	0.9	0.6	0.9	1.6	0.9	2.3	0.9	3.7	1.2	0.2	0.8	0.9
I	17.6	14.1	16.0	17.0	14.3	14.1	11.1	20.5	16.6	17.8	18.8	11.6	9.3	13.9	5.8	15.4	8.4	17.7	15.9	15.7
K	0.2	0.5	0.2	4.8	0.3	0.2	0.2	0.2	0.2	0.4	0.2	0.2	0.2	2.3	2.9	0.6	0.2	0.2	0.2	0.2
L	18.9	14.6	7.1	12.1	14.7	13.8	11.1	17.0	27.7	18.0	16.8	13.3	9.3	11.6	5.8	13.5	15.0	16.0	13.2	16.6
M	2.2	1.5	3.5	7.3	3.7	2.1	3.1	3.5	3.1	3.3	5.9	1.6	2.3	9.3	2.9	0.6	4.2	2.9	2.6	3.3
N	1.2	1.0	5.3	0.9	0.4	0.3	1.5	0.6	0.9	0.7	0.4	0.9	2.3	4.6	2.9	1.2	3.6	0.7	4.4	0.3
P	0.6	3.0	3.5	0.6	0.7	2.1	0.6	0.3	0.6	0.3	0.4	1.6	0.6	2.3	0.6	1.2	0.6	0.2	3.5	0.6
Q	0.4	0.5	0.6	0.6	0.9	0.3	1.5	0.5	5.5	0.4	1.9	3.3	2.3	0.6	2.9	1.8	0.6	0.6	1.7	0.3
R	0.2	0.5	3.5	0.5	0.4	0.3	0.5	0.1	5.5	0.1	0.4	1.6	0.5	2.3	0.5	0.6	2.4	0.6	1.7	1.5
S	2.2	3.5	8.9	4.8	3.6	1.8	9.5	2.3	5.5	2.1	0.4	3.3	4.6	6.9	2.9	8.6	0.6	1.6	4.4	2.1
T	2.9	3.5	1.7	2.4	2.9	3.6	3.1	1.2	2.5	2.4	3.4	10.0	2.3	2.3	11.7	0.6	2.4	3.0	1.7	3.3
V	22.2	15.1	21.4	4.8	19.3	22.1	6.3	22.2	16.6	21.1	19.8	16.6	6.9	20.9	26.4	13.5	25.3	25.4	12.3	19.7
W	1.8	3.5	1.7	1.7	1.5	1.0	1.5	1.6	1.7	1.4	1.4	8.3	9.3	4.6	5.8	3.0	1.2	1.0	3.5	2.7
Y	4.9	5.0	1.7	4.8	5.6	6.1	4.7	4.7	5.0	5.2	5.4	1.6	4.6	2.3	14.7	4.3	6.6	4.7	7.9	4.9

[a] The value in row i, column j is 100 * P (seeing residue i, given that it is aligned with residue j).

Table 2. Conditional Probabilities for Alignment of Exposed Residues from the β-Structure Database[a]

	A	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*K*	*L*	*M*	*N*	*P*	*Q*	*R*	*S*	*T*	*V*	*W*	*Y*
A	5.4	7.6	4.0	2.9	5.0	4.0	4.7	5.0	2.8	5.5	4.4	3.6	3.1	2.0	4.1	3.4	2.7	5.4	2.1	5.9
C	2.3	10.2	0.4	0.2	2.5	0.5	0.9	0.8	0.9	1.2	0.8	1.0	1.5	0.3	1.6	1.4	1.2	1.0	2.1	1.8
D	3.8	1.2	2.4	2.1	2.5	4.0	5.2	3.5	6.7	3.2	3.5	3.6	1.5	4.8	6.4	5.3	3.7	2.1	4.3	2.7
E	5.4	1.2	4.0	5.1	7.6	4.0	5.6	5.7	15.8	6.4	10.6	9.3	7.9	7.3	13.0	5.1	8.0	6.0	7.6	4.6
F	4.6	7.6	2.4	3.8	6.7	7.6	7.1	2.4	2.6	4.0	7.0	2.5	3.1	2.7	1.1	5.3	3.6	2.9	4.3	3.7
G	2.7	1.2	2.8	1.4	5.5	3.5	4.2	3.0	0.7	4.0	6.1	4.1	1.5	1.7	2.9	2.9	1.8	2.0	3.2	3.7
H	3.8	2.5	4.5	2.5	6.3	5.2	2.8	2.8	3.0	3.0	4.4	2.0	4.7	2.7	2.2	3.4	3.3	3.4	2.1	4.9
I	8.9	5.1	6.5	5.5	4.6	8.1	6.1	13.1	6.9	9.4	9.7	5.1	6.3	6.6	6.8	7.0	5.4	7.6	7.6	5.9
K	5.8	6.4	14.7	17.9	5.9	2.3	7.5	8.1	8.6	7.7	10.6	8.8	3.1	9.4	5.9	5.3	7.2	6.5	9.8	11.8
L	10.0	7.6	6.1	6.4	8.0	11.1	6.6	9.6	6.7	10.3	5.3	3.1	14.2	6.2	5.2	6.5	4.8	10.9	7.6	6.5
M	1.9	1.2	1.6	2.5	3.3	4.0	2.3	2.4	2.2	1.2	3.5	1.5	1.5	1.3	2.0	1.4	1.0	1.4	1.0	0.6
N	2.7	2.5	2.8	3.8	2.1	4.6	1.8	2.1	3.2	1.2	2.6	3.1	3.1	4.5	4.5	3.1	3.7	1.8	4.3	4.0
P	0.7	1.2	0.4	1.0	0.8	0.5	1.4	0.8	0.3	1.9	0.8	1.0	1.0	0.6	1.1	0.4	0.7	1.2	2.1	2.4
Q	2.3	1.2	5.7	4.4	3.3	2.9	3.7	4.1	5.0	3.8	3.5	6.7	3.1	8.3	3.4	3.9	5.9	4.5	6.5	4.6
R	6.9	8.9	11.4	12.2	2.1	7.6	4.7	6.5	4.9	4.9	7.9	10.3	7.9	5.2	2.7	6.8	8.4	8.2	10.9	5.9
S	5.4	7.6	9.0	4.4	9.3	7.0	6.6	6.3	4.1	5.7	5.3	6.7	3.1	5.5	6.4	8.7	10.3	4.5	5.4	6.8
T	6.9	10.2	10.2	11.3	10.1	7.0	10.4	7.9	9.0	6.8	6.1	12.9	7.9	13.6	12.8	16.5	15.7	11.3	2.1	4.3
V	11.6	7.6	4.9	7.0	6.7	6.4	9.0	9.2	6.7	12.8	7.0	5.1	11.1	8.7	10.3	6.0	9.3	11.3	9.8	9.3
W	0.7	2.5	1.6	1.4	1.6	1.7	0.9	1.5	1.6	1.5	0.8	2.0	3.1	2.0	2.2	1.2	0.3	1.6	2.1	1.2
Y	7.3	7.6	3.6	3.2	5.0	7.0	7.5	4.1	7.1	4.5	1.7	6.7	12.6	5.2	4.3	5.3	2.1	5.4	4.3	8.6

[a] The value in row i, column j is 100 * P (seeing residue i, given that it is aligned with residue j).

Furthermore, the correct alignment appears in the top five scoring alignments in 72 of the 77 pairs (as described below, the five top scoring aligned rungs are kept at each stage in generating the tree of wraps from an initial rung).

We remark that while the raw rung–rung score is based on non-β-helix 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-helix structures were included in the training set (for example, the gap penalty suffered in accuracy when 1TSP, which has a very long T1 loop, was not in the training set). The numeric values reported for the gap score above and for the α-helical content filter later in the manuscript are the ones for the final version of BetaWrap that incorporates all known structures.

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 five-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 (of -13) 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 (Fig. 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.* (1998), 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 permitted only 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 2, a penalty (of 2 minus the number of hydrophobics) 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 (Garnier *et al.*, 1996) was adapted to construct a two-state (α-with-high-confidence/other) secondary-structure predictor. We used the original GOR-IV training set and assigned an α-helical score to each residue exactly as detailed in Garnier *et al.* (1996). 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. A residue was considered a helix with high-probability only if its GOR-IV score exceeded -0.4. Call such a residue an "H" residue. A wrap was rejected if three or more of the rungs contained more than four "H" residues. Additionally, if greater than .33 of the residues across the full candidate wrap were "H" residues, the wrap was also rejected.

Sequences were prefiltered for trans-membrane α-helices using the GES hydrophobicity scale (Engelman *et al.*, 1996), a window of size 21, and a threshold of -2 kcal/mol. The predicted helices were removed, and the query sequence was broken into subsequences which were scored individually.

3. METHODS

3.1. The databases

The PDB-minus database was constructed from the PDB_select 25% list of June 2000 (Hobohn *et al.*, 1992; Hobohn and Sander, 1994), 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% sequence similarity.) The database contained 1,346 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 (Frishman and Argos, 1995), which annotates secondary structure, hydrogen bonds, and residue accessibilities. Those β-sheets whose residue accessibilities fit an alternating pattern of buried/exposed (relative accessibility [Rost and Sander, 1994]) alternating between < 0.05 and > 0.15) 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. Tables 1 and 2 present the scores for the pairwise amino acid stacking probabilities learned from this database.

New β-helices were identified from the sequence databases SWISS-PROT (Release 39.6 of August 30, 2000: 88,166 entries) and TrEMBL (Release 14.11 of August 25, 2000: 301,497 entries) (Bairoch and Apweiler, 2000).

3.2. Training

A seven-fold cross-validation was performed on the seven β-helix families of closely related proteins in the SCOP (Murzin *et al.*, 1995) database (see Footnote 1). PDB-minus was randomly partitioned into a 60% training (with 815 structures) and 40% testing (with 531 structures) set. For each cross, proteins in one β-helix family were placed in the test set, while the remainder of the β-helices were placed in the training set. The scores reported for the β-helix proteins in Table 3 and in Figure 4 are the scores in the leave-family-out cross experiment for that β-helix's protein family. The figure additionally reports the scores of all the PDB-minus proteins. 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 (Fig. 4). The score for each β-helix is

TABLE 3. KNOWN RIGHT-HANDED β-HELICES AND THEIR BETAWRAP SCORES/RANKS

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

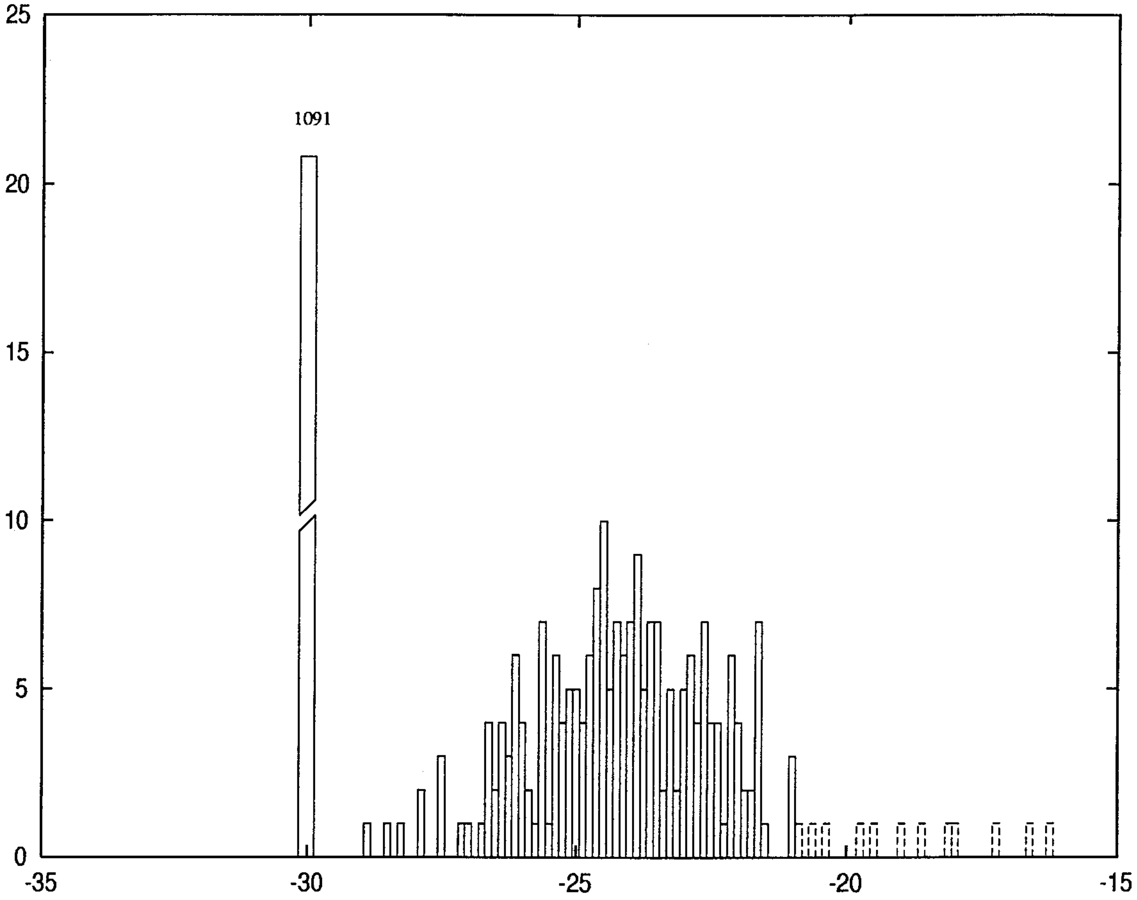

FIG. 4. Histogram of protein scores as computed by BetaWrap. The β-helix scores (12 proteins) were superimposed on the scores of the PDB-minus database (1,346 proteins), with the 1,091 proteins which could not be successfully wrapped (Section 2.2) given the arbitrary score −30. The β-helix histogram is dashed, and PDB-minus is solid.

taken from its cross-validation run. In Table 3, 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 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 12 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 4 of the 10 scored parses correct and the other 6 off in a single rung.

4.2. New β-helix candidates

The BetaWrap program has identified many new sequences that we believe contain β-helix structures. Table 4 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.* (1996) based on sequence similarity to P.69 pertactin. In the general list of the high-scoring proteins, the pectate lyases and galacturonases are well represented, as are the pollen allergens which are members of the pectate lyase superfamily and have been predicted to have a β-helical structure (Yoder *et al.*, 1993; Henrissat *et al.*, 1995). A significant fraction of the proteins found are characterized as outer membrane or cell-surface proteins; a significant number have roles in bacterial pathogenesis. For a more complete list, see *theory.lcs.mit.edu/betawrap.*

TABLE 4. EXAMPLES OF PROTEINS PREDICTED TO FORM β-HELICES BY BETAWRAP WITH THEIR SCORES[a]

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

[a] Identifiers (ID) and descriptions are taken from SWISS-PROT or TrEMBL. For an updated list of proteins with high BetaWrap scores, see *theory.lcs.mit.edu/betawrap.*

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-helices: PSI-BLAST (Altschul *et al.*, 1990), Threader (Jones *et al.*, 1992), and HMMer (Eddy, 1998).

First, the sequences of the 12 β-helix domains were used to search the NCBI nonredundant database (December 14, 2000, update, 595,890 entries) using the iterative multiple sequence alignment program PSI-BLAST (Altschul *et al.*, 1990) (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 5).

Next, the program Threader 2.5 (Jones *et al.*, 1992) was used to thread the 12 β-helix sequences onto an accompanying fold library (March 1999 version, 1,906 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.

Finally, we looked at the HMMer hidden Markov model program (hmmbuild from HMMER 2.1.1 [Eddy, 1998]). The input to HMMer is a multiple sequence alignment. We remark that, because the best multiple sequence alignments are typically constructed using full information as to how the structures align in 3D,

Table 5. Results of PSI-BLAST Searches on the Known β-Helix Structures[a]

	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

[a] 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.

Table 6. Results of HMMer Searches with the Rows Representing the Initial Seed Aligned by FSSP (One Structure Chosen from Each β-Helix Family), and the Columns, All the β-Helix Structures Indexed by Their PDB Codes (see Table 3)[a]

	1PLU	*1PCL*	*1BN8*	*1IDK*	*1QCX*	*1RMG*	*1BHE*	*1CZF*	*1TSP*	*1DAB*	1DBO	*1QJV*
F1	—	—	—				11	19	2	1	3	
F2				—	—		21	4	2	12		
F3	1	6	2		1	—	—	—	2	3	1	5
F4		3	1				5	11	—	1		4
F5	52	74	73	16	31		49	22	2	—	1	99
F6	1		1				14	2	7	4	—	1
F7							28	3	6	7	4	—

[a] The numerical entries count the number of non-β-helix sequences that the HMM scored above the β-helix sequence, when its family was left out of the multiple sequence alignment fed to HMMer. A dash indicates a protein from the same family as the seed.

this is not a strictly sequence-based method. In spite of this advantage, we find it does not perform as well as BetaWrap. We used the standard structural alignment program, FSSP (Holm and Sander, 1996), to construct seven different multiple sequence alignments, each one based on a seed from the FSSP data for a structure in a different β-helix family (see Table 6). The particular seeds to FSSP were F1:1PCL, F2:1QCX, F3:1BHE, F4:1TSP, F5:1DAB, F6:1DBO, F7:1QJV. Each of these initial alignments was then used to do a separate seven-fold leave-one-out cross-validation, where the multiple alignment for the HMM is built using each seed from the six families in the training set (so Table 6 represents 6×7 cross validation experiments). Note that a dash in the table indicates that the family left out of the cross validation was the same family as the FSSP seed. As can be seen, each one of the seven seeded HMM models fails to separate positive and negative examples on PDB-minus. In terms of overlap of positive and negative score values, the F3 seed gets the fewest false positives when the threshold is set to find all beta-helices, and the F3 seed has the fewest false negatives when the threshold is set to not have false positives. However, for all seven, four β-helix structures are universally missed (1BHE, 1CZF, 1TSP, and 1DAB).

6. 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 3) 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 [Sonnhamer *et al.*, 1998]). 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, Berger and Singh (1997).

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 (Haase-Pettingell and King, 1997).

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.

7. BETAWRAP ON THE WEB

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

ACKNOWLEDGMENTS

This work was supported in part by the Charles E. Reed Faculty Initiatives Fund. L.C. was a visitor at MIT supported by an Emmaline Bigelow Conland Fellowship at the Radcliffe Institute for Advanced Study. J.K. was supported in part by NIH GM 17980. P.B. was supported by an MIT/Merck graduate fellowship. B.B. was supported in part by an NSF Career Award. Thanks to Scott Decatur, Jonathan Dunagan, Brian Dean, and Dan Kleitman for their initial computational assistance; to Robert Seckler for suggesting training sequences during the development of the algorithm; 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.

REFERENCES

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

Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W., and Lipman, L.. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucl. Acids Res.* 25, 3389–3402.

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

Berger, B. 1995. Algorithms for protein structural motif recognition. *J. Comp. Biol.* 2, 125–138.

Berger, B., and Singh, M. 1997. An iterative method for improved protein structural motif recognition. *J. Comp. Biol.* 4(3), 261–273.

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

Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I., and Bourne, P. 2000. The Protein Data Bank. *Nucl. Acids Res.* 28, 235–242.

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

Eddy, S. 1998. Profile hidden Markov models. *Bioinformatics* 14, 755–763.

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

Engelman, D., Steitz, T., and Goldman, A. 1986. Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins [review]. *Ann. Rev. Biophysics Biophysical Chem.*, 15, 321–53.

Frishman, D., and Argos, P. 1995. Knowledge-based secondary structure assignment. *Proteins: Struct. Funct. Genet.* 556–579.

Garnier, J., Gibrat, J., and Robson, B. 1996. GOR secondary structure prediction method version IV. *Methods Enzymol.* 266, 540–553.

Haase-Pettingell, C., and King, J. 1997. Prevalence of temperature sensitive folding mutations in the parallel beta coil domain of the phage P22 tailspike endorhamnosidase. *J. Mol. Biol.* 267, 88–102.

Heffron, S., Moe, G., Sieber, V., Mengaud, J., Cossart, P., Vitali, J., and Jurnak, F. 1998. Sequence profile of the parallel β helix in the Pectate Lyase superfamily. *Struct. Biol.* 122, 223–235.

Henrissat, B., Heffron, S., Yoder, M., Lietzke, S., and Jurnak, F. 1995. Functional implications of structure-based sequence alignment of proteins in the extracellular pectate lyase superfamily. *Plant Physiol.* 107, 963–976.

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

Hobohm, U., Scharf, M., Schneider, R., and Sander, C. 1992. Selection of a representative set of structures from the Brookhaven Protein Data Bank. *Protein Sci.* 1, 409–417.

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

Huang, W., Matte, A., Li, Y., Kim, Y., Linhardt, R., Su, H., and Cygler, M. 1999. Crystal structure of chondroitinase B from *Flavobacterium heparinum* and its complex with a disaccharide product at 1.7 Å resolution. *J. Mol. Biol.* 294, 1257.

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

Jenkins, J., Mayans, O., and Pickersgill, R. 1998. Structure and evolution of parallel β-helix proteins. *Struct. Biol.* 122, 236–246.

Jenkins, J., Mayans, O., Smith, D., Worboys, K., and Pickersgill, R. 2001. Three-dimensional structure of *Erwina chrysanthemi* pectin methylesterase reveals a novel esterase activity site. *J. Mol. Biol.* 305, 951.

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

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

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

Karplus, K., Barrett, C., and Hughey, R. 1998. Hidden Markov models for detecting remote protein homologies. *Bioinformatics* 14, 846–856.

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

Kreisberg, J., Betts, S., and King, J. 2000. Beta-helix core packing within the triple-stranded oligomerization domain of the P22 tailspike. *Protein Sci.* 9, 2338–2343.

Lifson, S., and Sander, C. 1980. Specific recognition in the tertiary structure of β-sheets of proteins. *J. Mol. Biol.* 139, 627–629.

Mayans, O., Scott, M., Connerton, I., Gravesen, T., Benen, J., Visser, J., Pickersgill, R., and Jenkins, J. 1997. Two crystal structures of pectin lyase A from *Aspergillus* reveal a pH driven conformational change and striking divergence in the substrate-binding clefts of pectin and pectate lyases. *Structure* 5, 677.

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

Petersen, T., Kauppinen, S., and Larsen, S. 1997. The crystal structure of rhammogalacturonase A from *Aspergillus aculeatus*: A right-handed parallel beta helix. *Structure* 5, 533.

Pickersgill, R., Jenkins, J., Harris, G., Nasser, W., and Robert-Baudouy, J. 1994. The structure of *Bacillus subtilis* pectate lyase in complex with calcium. *Nat. Struct. Biol.* 1, 717.

Pickersgill, R., Smith, D., Worboys, K., and Jenkins, J. 1998. Crystal structure of polygalacturonase from *Erwina carotovora* ssp. carotovora. *J. Biol. Chem.* 273, 24660–24664.

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

Rost, B., and Sander, C. 1994. Conservation and prediction of solvent accessibility in protein families. *Proteins* 20(3), 216–26.

Singh, M., Berger, B., and Kim, P.S. 1999. Learncoil–VMF: Computational evidence for coiled coil-like motifs in many viral membrane fusion proteins. *J. Mol. Biol.* 290(1), 241–251.

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

Sippl, M., and Weitckus, S. 1992. 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.

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

Steinbacher, S., Sekler, R., Miller, S., Steipe, B., Huber, R., and Reinemer, P. 1994. Crystal structure of P22 tailspike protein: Interdigitated subunits in a thermostable trimer. *Science* 265, 383.

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

VanSanten, Y., Benen, J., Schroter, K., Kalk, K., Armand, S., Visser, J., and Dijkstra, B. 1999. 1.68 Å crystal structure of endopolygalacturonase ii from *Aspergillus niger* and identification of active site residues by site-directed mutagenesis. *J. Biol. Chem.* 274, 30474.

Vitali, J., Schick, B., Kester, H., Visser, J., and Jurnak, F. 1998. The three-dimensional structure of *Aspergillus niger* pectin lyase B at 1.7-å resolution. *Plant Physiol.* 116, 69.

Wolf, E., Kim, P.S., and Berger, B. 1997. `MultiCoil:` A program for predicting two and three stranded coiled coils. *Protein Sci.* 6(6), 1179–1189.

Yoder, M., and Jurnak, F. 1995. The refined three-dimensional structure of pectate lyase C from *Erwinia chrysanthemi* at 2.2 angstrom resolution. *Plant Physiol.* 107, 349.

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

Yoder, M.D., and Jurnak, F. 1995. Protein motifs. 3. The parallel beta helix and other coiled folds. *FASEB J.* 9(5), 335–42.

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

Zhu, H., and Braun, W. 1999. 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.

Address correspondence to:
Dr. Bonnie Berger
Department of Mathematics
Massachusetts Institute of Technology
Cambridge, MA 02139

E-mail: bab@mit.edu