Markov random fields reveal an N-terminal double beta-propeller motif as part of a bacterial hybrid two-component sensor system

Matt Menke[a,b], Bonnie Berger[b,1], and Lenore Cowen[a,1]

[a]Tufts University, Medford, MA 02155; and [b]Massachusetts Institute of Technology, Cambridge, MA 02139

Edited by Peter W. Shor, Massachusetts Institute of Technology, Cambridge, MA, and approved November 30, 2009 (received for review August 31, 2009)

The recent explosion in newly sequenced bacterial genomes is outpacing the capacity of researchers to try to assign functional annotation to all the new proteins. Hence, computational methods that can help predict structural motifs provide increasingly important clues in helping to determine how these proteins might function. We introduce a Markov Random Field approach tailored for recognizing proteins that fold into mainly β-structural motifs, and apply it to build recognizers for the β-propeller shapes. As an application, we identify a potential class of hybrid two-component sensor proteins, that we predict contain a double-propeller domain.

remote homology detection | motif recognition | structure | signal transduction | histidine kinase

Bacteria are adept at sensing and adjusting to conditions in their environment. For a bacteria to respond to its environment, it needs to be able to monitor extracellular changes including osmotic activity, ionic strength, pH, temperature, and the concentrations of nutrients and harmful compounds (1). Frequently, such processes are mediated by two-component sensor proteins, involving a usually membrane-bound sensor protein, and a response regulator. A set of 32 unusual hybrid periplasmic-sensing two-component sensor histidine kinases were discovered in the human gut symbiont *Bacteroides thetaiotaomicron* (2). These hybrid two-component sensor systems (HTCS) were unusual in that they incorporated all of the domains found in the classical two-component system into a single polypeptide (3). Sonnenburg et al. (2) hypothesized that the HTCS proteins were involved in how these microbiota sense diverse nutrients and implement an appropriate metabolic response, and showed that loss of one of these HTCS proteins, BT3172, reduced glycotic pathway activity. They also found that the sensor domains of the HTCS proteins were less highly conserved than their intracellular signaling domains, suggesting that the HTCS proteins have diversified to respond to a variety of signals while conserving their means of intracellular signal transduction. HTCS proteins have since been found in other prokaryote genomes, primarily in Bacteroidetes and Proteobacteria (1, 2).

In this article, we use computational methods to predict that over 300 bacterial proteins (primarily from Proteobacteria and Bacteroidetes) have a unusual double β-propeller motif in their N-terminal region, followed by Pfam's two-component regulator three Y (YYY) motif (4), followed by most commonly a histidine kinase domain [but also, sometimes a diguanylate cyclase domain (GGDEF) or a stage 2 sporulation E protein (SpoIIE) domain, and others], see Fig. 1. Many of the histidine kinase domain-containing sequences also have a response regulator signature as well, confirming their role as hybrid two-component sensor systems. The prediction of the double-propeller motif was accomplished by use of a unique computational method that employs Markov random fields to predict β-structural motifs in distantly homologous proteins. Weak sequence homology already suggested that the N-terminal transmembrane sensing region contained

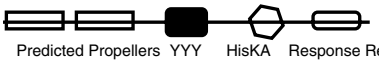

Fig. 1. Schematic of a hybrid two-component sensor YYY protein with a histidine kinase HisKA domain; we predict two beta propellers in the N-terminal region.

propeller blades (see below), but could not have determined their exact number or how they split into different propellers, though Mascher et al. (1) suggested that this region of the HTCS histidine kinases in *B. thetaiotaomicron* could form two seven-bladed propellers.

Markov random field (MRF) methods (5) that generalize hidden Markov models (HMM) can allow arbitrary dependencies between nonadjacent states, to better model β-structural motifs such as β-propellers in protein sequence. Using an MRF framework, it is possible to model interactions between residues that participate together in a secondary or super-secondary structural motif, *but may not be close together in the 1D sequence*. Such an approach should particularly assist in the prediction of protein structural motifs whose secondary structure is "mainly beta." This is because although there is evidence that residues involved in β-sheet formation that are close in space exhibit strong statistical biases (6–8), these residues may be difficult to discover due to being a variable and potentially long distance apart in the protein sequence. In fact, simply predicting the correct annotation of secondary structure of these folds can be problematic: Even the best secondary structure predictors such as PHD (9) and PSIPRED (10) predict α-helices more accurately than β-strands (11). Tertiary structure predictors such as Rosetta (12) and LINUS (13), although performing well on all-α- and α/β-proteins, are also challenged by topologically complex all-β-proteins (12, 13). Many threading programs also have particular problems recognizing and then threading β-sheet topologies correctly once sequences fall into the so-called twilight zone (14) of less than 15–20% sequence homology to known structures.

Previous methods have been introduced by our group and others to identify β-structural motifs from sequence by capturing pairwise dependencies between residues that come together to form the β-sheets of the motif (8,15–19). These methods were shown to be more successful than a variety of competing methods at recognizing the right-handed parallel β-helix fold (8,15–18), the β-trefoil fold (18, 19), TM-barrel proteins (20), and Leucine-Rich repeat folds (16). These methods have been shown to

Author contributions: B.B. and L.C. designed research; M.M. and L.C. performed research; M.M. contributed new reagents/analytic tools; M.M., B.B., and L.C. analyzed data; and M.M., B.B., and L.C. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1]To whom correspondence may be addressed. E-mail: bab@mit.edu or cowen@cs.tufts.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0909950107/DCSupplemental.

produce meaningful results on particular β-structural motifs *despite the fact that they are throwing away almost all sequence information (of the type learned by a profile HMM)*. None of the previous methods solved the elusive problem of combining both HMM information and higher-order β-sheet dependency information at the same time.

In this article, we show how to combine both HMM information and higher-order β-sheet dependency information into a single, integrated MRF whose energy function can be solved exactly in a computationally tractable way using multidimensional dynamic programming. We introduce the structural motifs using random fields (SMURF) framework and apply it to the recognition of the large classes of β-propeller folds. We test it on solved protein structures comprising three structural motifs in stringent leave-family-out cross-validation experiments; namely the six-, seven-, and eight-bladed β-propellers. These β-propellers are ideal structural motifs on which to test our methods because they are highly diverse in sequence, taxa, and function (21), with some families having a strong sequence signature and others not, while also having a repeating β-strand topology (see Fig. 2). We then chain two six-, seven-, or eight-bladed SMURF propeller models together to construct a recognizer of a double-propeller motif that we test on the putative hybrid two-component regulatory system proteins we discuss above, with strong positive results. We show SMURF is better able to recognize this potential motif than HMM-based models on the one hand, and also better than an MRF that measures pairwise propensities for β-sheet formation alone.

To test the double-propeller hypothesis, we constructed nine SMURF-based templates of two β-propellers: namely, two six-bladed, a six-bladed followed by a seven-bladed, a six-bladed followed by an eight-bladed, etc. Using a SMURF-based template of two seven-bladed propellers, we now find over 300 proteins predicted to form double propellers, followed by a YYY motif, in over 100 bacterial species, primarily in Proteobacteria and Bacteroidetes (see *Results* and *SI Text*). We show that a domain architecture of twin seven-bladed propellers is computationally favored by the most structures over six-bladed, eight-bladed, or any mix of the other numbers of blades in the propellers, and no structures are predicted to contain eight-bladed propellers. Based on sequence motifs that follow the YYY motif in sequence, these proteins appear to also be two-component regulatory system proteins.

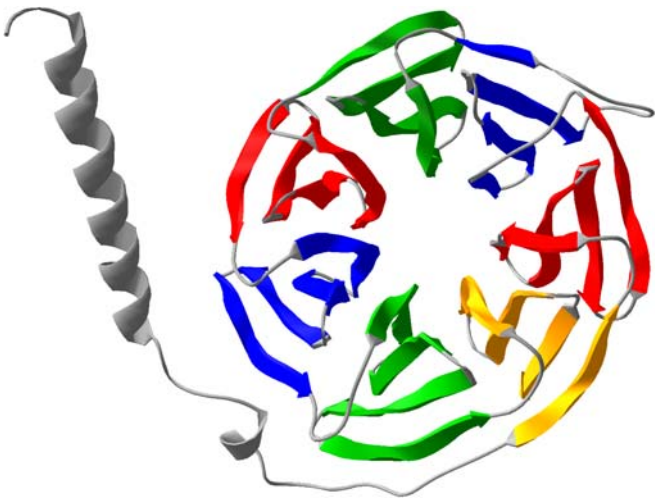

Fig. 2. Seven-bladed β-propeller.

Results

Predicting Propeller Folds. We tested our methods on the superfamilies of the six-, seven-, and eight-bladed propeller folds, where we performed leave-family-out cross-validation on each one. We compared this to the latest version of one of the most popular HMM programs, HMMER 3.0 (22). (For both SMURF and HMMER, we removed all six-, seven-, and eight-bladed propellers from the negative set; distinguishing the number of blades in the propeller is a more difficult problem that we discuss below.) Recall that SMURF has a two-component score, one component that roughly corresponds to HMMER alone, and one that measures pairwise correlations, which it gives equal weight. As can be seen in Table 1, SMURF outperforms HMMER in every category, often substantially. We also wanted to get a sense of how powerful the pairwise probabilities alone would be, so the SMURF(P) column in the table shows the results of our SMURF algorithm reweighted so that the HMM component score gets a weight of zero and the pairwise correlation score gets a weight of one. From Table 1, it can be seen that this performs substantially worse than SMURF or even HMMER on the six-bladed and seven-bladed propellers. Strangely, however, SMURF(P) performs best of all three methods on the eight-bladed propellers. The eight-bladed propellers are the fold class where there are the fewest training examples, and thus it is not surprising that the HMM/sequence-based component is least helpful in cross-validation in this case. This suggests that for each new structural motif, the two components of the SMURF score should be weighted according to how good the training data are for a sequence-based model. This weighting can be part of the model building for SMURF (we did not do it here because we wanted to make sure we did not overtrain SMURF when reporting cross-validation results).

We also ran SMURF on a large database of sequences of unsolved structures, namely, version 14.9 of Uniprot filtered to 50% sequence identity (23). Lists of the top 1,000 proteins that SMURF predicts to form six- or seven-bladed propellers can be found in the *SI Text*. A webserver that will accept any protein sequence and score the likelihood that it matches each propeller template is available at http://smurf.cs.tufts.edu.

Two-Component Sensor YYY Proteins. Pfam (4), a database of protein sequence motifs derived from HMMER models, identifies 506 protein sequences, putative protein sequences, or fragments as containing instances of a YYY motif. Of these, 237 are predicted to contain the HisKA motif by Pfam, signature of a histidine kinase domain. We built nine double-bladed β-propeller templates by simply chaining two six-, seven-, or eight-bladed SMURF propeller templates together, in all combinations. For each template, we compute the best scoring match to the pair of propellers using dynamic programming. In the 506 YYY-motif containing proteins, Pfam predicts 102 to have no instances of the motif RegProp, and between 3 and 13 instances of the motif RegProp in the remaining proteins, where RegProp is supposed (by sequence similarity) to be homologous to a blade of a propeller. Restricted to proteins that contain both the YYY and the HisKA motif, Pfam also predicts between 0 and 13 instances of the motif RegProp. Fig. 3 shows the number of instances of the motif predicted by Pfam in the N-terminal region of proteins where both a YYY and a HisKA motif is recognized.

We constructed nine templates, based on chaining together two of the SMURF six-,seven-, and eight-bladed propeller templates, consisting of (*a*) two six-bladed propellers, (*b*) a six-bladed followed by a seven-bladed propeller, (*c*) a seven-bladed followed by a six-bladed propeller, (*d*) a six-bladed followed by an eight-bladed propeller, etc. Of the 506 YYY-motif proteins, there were 475 for which at least one template was accepted with a p value <0.01 for containing the motif. Table 2 breaks down the SMURF predictions of the double-propeller motif by p value; we have high confidence the motif is present when the p value is less than 0.001;

Table 1. Our results versus HMMER 3.0

	Six-bladed			Seven-bladed			Eight-bladed		
TN	HMMER	SMURF(P)	SMURF	HMMER	SMURF(P)	SMURF	HMMER	SMURF(P)	SMURF
97%	52	20	**80**	80	23	**87**	0	**40**	0
96%	56	24	**80**	80	33	**87**	20	**40**	**40**
95%	64	36	**80**	87	47	**93**	20	**40**	**40**
94%	68	36	**84**	90	53	**93**	40	**60**	40
93%	68	48	**84**	90	53	**97**	40	**100**	40
92%	68	60	**88**	90	57	**97**	40	**100**	40
91%	68	60	**92**	90	57	**97**	40	**100**	40
90%	68	60	**92**	93	57	**100**	60	**100**	**100**

The numbers represent the percent of true positives correct for a given threshold of percent of true negatives (TN) on a leave-superfamily-out cross-validation. Best results (in bold) come from SMURF, or, on the eight-bladed propellers, SMURF(P) which reweights SMURF away from the HMM entirely to consider only our pairwise scores. Note that structures with fewer than 150 residues were removed from the test set, as they are too short to fold into β-propellers with six or more blades.

we predict the motif (but with lower confidence) when the *p* value is between 0.001 and 0.01. Many of these sequences appear to be two-component regulators, with the propeller domain preceding the YYY in sequence, and a known sensing domain (such as a histidine kinase or the GGDEF domain) appearing after the YYY. We also list SMURF scores for the three most common families by signature domain after the YYY motif: the HisKA histidine kinase domain, the GGDEF adenylyl cyclase signaling domain, and the SpoIIE sporulation domain. (Other domains that Pfam recognizes in at least one sequence that appear after the YYY domain include cGMP phosphodiesterase/adenylcyclase/Fhla (Gaf), Per/Art/Sim (PAS), and Histidine kinase-like ATPase (HATPase) domains.) Many of the histidine kinase structures also have a detectable "response regulator" sequence signature after the histidine kinase, confirming their role as two-component sensor systems. SMURF scores a double propeller with confidence in a greater proportion of the HisKA containing protein sequences than in the GGDEF or SpoIIE domain-containing protein sequences. We note that 24 of the lowest-scoring proteins are clearly fragments. An additional 26 of the low-scoring proteins consist of proteins with multiple copies of the YYY domain that do not appear to be two-component regulators. Their conjectured function remains unknown.

In terms of the species distribution, SMURF tends to score highest YYY proteins that come from the Bacteroidetes, particularly the Bacteroidacea subclass, and nearly all the highest-confidence double-propeller predictions come from *Bacteroides* species. Proteins from the *Xanthomonas* genus tend to be lower scoring, indicating perhaps either a less common sequence pattern (which would lower the HMM score component), or a less regular propeller blade structure (which would lower the pairwise score component). Three figures describing the species distribution of the YYY sequences appear in *SI Text*.

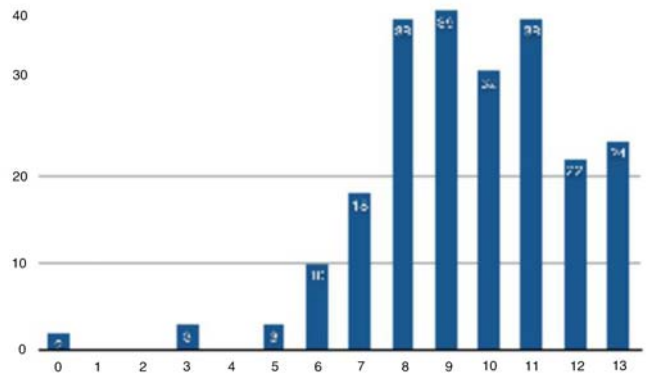

Fig. 3. The *x* axis of this histogram gives the number of "blade" motifs that Pfam predicts in the N-terminal region, and the *y* axis gives the number of structures for which Pfam predicts that number of blades. Pfam (based on an HMM alone) fails to predict a double-propeller structure, instead predicting between 0 and 13 propeller blades in the N-terminal region of the 237 protein sequences containing both a YYY and a HisKA domain. SMURF predicts these blades are organized into two β-propellers.

Because propellers of differing numbers of blades appear to share considerable homology (24), like other homology-based methods such as HMMs, SMURF is better at distinguishing propellers from nonpropellers than in determining the number of blades, particularly because the seventh blade is so irregular that it is not consistently captured in the SMURF template. For our original templates, we found that, although SMURF predicts them to be less likely to contain eight-bladed propellers, it could not distinguish between 6–6, 6–7, 7–6, or 7–7 templates reliably. To address this issue, we built a SMURF seven-bladed propeller template from only 23 of the 30 solved seven-bladed propeller structures in the nonredundant protein data bank (PDB), excluding the seven-bladed propellers that had the most irregularity between blades six and seven (namely, the seven-bladed propellers in 1mda, 1a0r, 2i3s, 1u4c, 1jtd, 1utc, and 1c9i). This unique seven-bladed template scores all 30 seven-bladed propellers higher than 75% of the six-bladed propellers in the nonredundant PDB in stringent cross-validation, whereas the original seven-bladed template scores both six- and seven-bladed propellers highly. Substituting in the new, stricter seven-bladed template, we created four new double-propeller templates: two six-bladed, a six-bladed followed by a seven-bladed, a seven-bladed followed by a six-bladed, and two seven-bladed. We considered the 478 of the 506 YYY structures that have a *p* value of 0.01 or better for one of those four templates. For each template, the percentage of the YYY-containing sequences on which it scores best appears in Table 3. If we conjecture that the motif is conserved across all the YYY structures that contain double propellers, we thus predict that the motif consists of two seven-bladed propellers. The full list

Table 2. SMURF double-propeller predictions for the proteins containing Pfam's YYY motif

Motif	Total	<0.0001	<0.001	<0.01	≥0.01
all	506	80	259	136	31
HisKA	237	59	143	29	6
GGDEF	90	9	36	42	3
SpoIIE	22	0	16	6	0

The columns indicate the SMURF *p* value for the best of the nine templates. In addition to statistics for all YYY sequences, we also report separately performance on several large classes for which there is a sequence motif signature after the YYY: the HisKA family of Histidine kinases, the GGDEF family, and the SPoIIE family.

Table 3. Percentage of predicted double-bladed propellers

Blade number	6	7
6	18.8	14.4
7	25.7	**41.0**

Each row represents the number of blades in the first propeller and each column the number of blades in the second propeller for each of the four double-propeller templates. The entry for each template is the percent of the time that template has the lowest SMURF score (over the 478 proteins in which SMURF predicts a double-propeller with p value at most 0.01).

of SMURF double seven-bladed propeller predictions appears in *SI Text*.

Discussion

A major challenge in protein structural motif recognition has been to find a tractable way to integrate local sequence information with higher-order structural dependencies in an integrated, computationally tractable, energy function. We have presented a framework, SMURF, that integrates long-distance pairwise correlations involving β-sheet formation with an HMM using a random field, and tested it on three fold classes, namely, the five-, six-, and seven-bladed propellers. There is nothing in the method, however, that restricts it to propellers. The method can be applied to any β-structural motif where there is enough training data, and the number of interleaved pairs of β-strands is not too great (to maintain feasible computational resources). We will be applying SMURF to additional classes of β-structural motifs in the future.

We predict that the HTCS proteins containing the Pfam YYY motif contain double β-propellers; however, it seems that predicting propeller motifs is much easier than determining the number of propeller blades. As can be seen in Table 3, the template that is preferred most often is the twin seven-bladed propeller, over 40% of the time, with the next most preferred template the seven-bladed followed by the six-bladed propeller, over 25% of the time. Importantly, on a structure-by-structure basis, typically, even when the 7–7-bladed template is not preferred, it scores nearly as well as the most preferred template (*SI Text*). In cross-validation, the seven-bladed propellers scored well on both six- and seven-bladed templates, whereas the true six-bladed propellers scored poorly on seven-bladed templates. This is strong evidence for the presence of a seventh blade.

It is possible that some of the YYY structures fold into two seven-bladed propellers and some fold into a seven-bladed followed by a six-bladed propeller, etc. On the other hand, they might all fold into twin seven-bladed propeller structures, but the difficulty in locating the last blade of the seven-bladed propeller arises from the known issues of irregular "closures" in known propeller motifs. In particular, different instances of β-propeller motifs have different topologies by which they "close" the propeller between the N and C terminal ends of the structure (25, 26). For example, although most sequences in our dataset get very comparable scores on the four templates, the protein A7GIL1 and its close homologs from *Clostridium botulinum* all strongly prefer the 6–6 template to any of the templates with a seven-bladed propeller. Perhaps this structure has an atypical seventh blade or has lost the seventh blade entirely.

SMURF is a purely statistical method that is more powerful than an HMM in predicting β-structural motifs. To improve discrimination between six- and seven-bladed propeller templates, we will likely have to move beyond statistical modeling of the sequence to geometric modeling using, for example, sidechain packing (27, 31).

Methods

Fig. 4 shows how a profile HMM is typically designed for protein motif recognition. In addition to start and end states, there are match states corresponding to each sequence position, insertion states, and deletion states. In SMURF, the main difference is that we do not allow insertions or deletions in positions corresponding to match states along a β-strand: that is, once we match a residue in sequence with the first residue in a β-strand, subsequent residues must match to the following β-strand positions until the end of the β-strand without insert or delete states. To run SMURF, it is first necessary to parse the match states of the HMM to indicate which match states participate in β-strands, and which tuples of β-strand residues are to be paired. Below, we show how to generate such a parse automatically from a protein multiple-structure alignment.

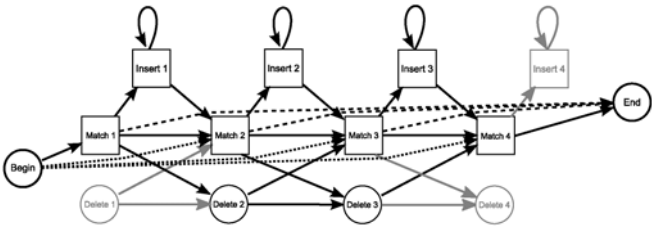

Fig. 4. States of a profile HMM.

The dependency diagram for SMURF for a pair of adjacent antiparallel β-strands appears in Fig. 5. An edge between two states indicates that the probabilities for those states are dependent on one another. The way we approximate the scores of the MRF is to decompose our log-likelihood score into a linear combination of two components. The first component is an HMM component, and the second component is a β-strand pair component. The HMM component is calculated as if the long-range dependency edges were not there; that is, it is just the score that the profile HMM would give for placing a residue in that match state. The β-strand pair component is calculated using pairwise probability frequency tables similar to our previous program Betawrap (8, 15); the only difference is that true propeller structures are removed from the training set to avoid overtraining for the cross-validation experiment. Here the pairs indicate the likelihood that two residues prefer to be hydrogen-bonded in a β-sheet. The formal transition probabilities are described in the next section.

Template Construction. For each of the known 3D structures from the fold class [according to Structural Classification of Proteins (SCOP) version 1.73] in the training set, it is marked which residue positions participate in a β-strand (using the Rasmol algorithm to decide if a position participates in a beta strand; see ref. 28). Each residue in a β-strand also determines which residues in which other β-strand it is paired with using the same program. The training structures are then aligned using the Matt multiple-structure alignment program (29), and we call a position in the alignment β-*conserved* if more than half the structures in the alignment mark that position as participating in a β-sheet. A pair of positions in the alignment is a β-*conserved pair* if more than half the structures in the alignment mark both positions as being hydrogen bonded with each other in the β-sheet. Two beta-conserved pairs AB and CD are said to be *adjacent* if in more than half the structures of the alignment, A is adjacent to C and B is adjacent to D, and AB and CD hydrogen bond to each other in the same β-sheet. Template β-conserved strands consist of the maximal contiguous sets of adjacent β-conserved strand pairs, together with the information of which residue positions are hydrogen bonded. Note that, because these template positions may only be occupied by positions that are identified as beta-strands in a majority of the structures, there could be residues in other structures between these positions in the Matt alignment, creating gap positions in the structural alignment in the middle of the template β-conserved strands. This is undesirable, so we will

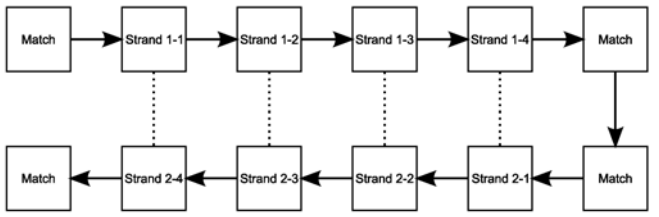

Fig. 5. SMURF states with pairwise dependencies (dotted lines) for two hydrogen-bonded antiparallel β-strands, each four residues long, with two intervening match states. Note there would also be three insertion and two deletion states between the β-strands.

remove these positions, *which includes possibly deleting residues from sequences when we remove the corresponding position from the template structural alignment*. In the resulting alignment, the beta-conserved template strands are always contiguous (see Fig. 6).

The output of this phase is a sequence alignment derived from the Matt structural alignment with annotated β-strand residue positions and the residue positions to which they are paired. This multiple sequence alignment, together with the locations of the template β-strands is the input to the next (training) phase.

The training phase trains both an HMM on the multiple sequence alignment, and also pairwise probabilities on the paired residue positions in the template beta strands. The HMM portion of the training is exactly the default HMM training from HMMER version 3, except that every template β-strand position is included in the output model. For the pairwise probabilities in the template β-strands, the "in" and "out" pairwise probability tables similar to the program Betawrap (8) are considered, but updated to reflect probabilities in the current nonredundant PDB (nrpdb.120707) version, and with all β-propeller protein structures (according to SCOP 1.73) removed. It is determined for each residue position, whether the sum of the pairwise probabilities of seeing the pairs in the training alignment, summed over all structures in the alignment, was more probable from the "buried" or the "exposed" probability table. The position is then labeled buried or exposed accordingly. Note that, although the names buried and exposed come from the solvent accessibility of residue positions in the original Betawrap algorithm, no claim about solvent accessibility or consistency of solvent accessibility is made in our context. In a structurally oblivious way, we are simply determining based on the training data which of the two models best predicts the set of residue pairs we see. The output is an HMM where certain states are marked as β-strand states, and pairings of hydrogen-bonded β-strand states are also given. Because the score includes a component from these pairs, and not just the linear score of the HMM component, it is no longer a simple HMM, but rather an MRF, because of the nonlinear dependencies between the paired states in constructing the score.

For a given test sequence, we seek the parse of the sequence to the states of the training HMM that optimizes the particular score. The score we seek to optimize, over all ways to map the sequence to the states of the MRF model, is

$$\alpha \times \log(\text{HMM score}) + \gamma \times \log(\text{pairwise score})$$

where the HMM score represents the conditional probability of seeing the sequence given the HMM portion of the model. Experimentally, using a value of one for both α and γ provided the best results on cross-validation.

We now specify the exact form of the transition probabilities for the MRF. Let the sequence have residues $r_1, \ldots, r_n$, and the MRF have match states $m_1, \ldots, m_l$, deletion states $d_1, \ldots, d_l$, and insertion states $i_1, \ldots, i_{l-1}$. And suppose that $r_1 \ldots r_k$ and match states $m_1 \ldots m_s$ have been assigned. Then, for the probability of assigning r_k to the next match state $m_j = m_{s+1}$ is

Before:

```
Seq1   ...VVDGD-ALLV--GFSEGSVN-YLYDG-GET-KLR--ING...
Seq2   ...VVDGDK--LLV-GFSEGSLQ-SMYDS-GETVKLR--ING...
Seq3   ...LD-GDLIA--FVS----RGQAFIQDSVGTYVL--KVL--...
Seq4   ...VI-GDL--IAFVS----RGY----DSVGTYVLKV--L--...
Seq5   ...VI-GDL--IAF-S----AGY--IQDSVGTY-LKV--L--...
              1  2345                   5432  1
```

After:

```
Seq1   ...VVDGD-  LV--GFSEGSVN-YLYDG-GET-KLR  ING...
Seq2   ...VVDGDK  LLV-GFSEGSLQ-SMYDS-GETVKLR  ING...
Seq3   ...LD-GDL  --FVS----RGQAFIQDSVGTYVL--  L--...
Seq4   ...VI-GDL  IAFVS----RGY----DSVGTYVLKV  L--...
Seq5   ...VI-GDL  IAF-S----AGY--IQDSVGTY-LKV  L--...
              1  234                    432  1
```

Fig. 6. Example of the template construction phase modifying a pair of hydrogen-bonded antiparallel β-strands in a Matt alignment containing five structures. The positions labeled with the same number are hydrogen-bonded to each other, in all sequences with a residue in corresponding numbered positions. Intervening residues are removed from the four-residue consensus β-strand pair. Fewer than half the structures have a residue in position five in the first strand, so both position fives are removed from the β-strand. Residues outside all consensus β-strand pairs are never removed.

$$\Pr[m_j|r_k, u_{j-1}] = \text{HMM}[m_j, r_k] \cdot \text{transition}[u_{j-1}, m_j] \cdot \beta\text{-strand}[r_j, r_k, m_j, m_k],$$

where u_{j-1} can be either d_{j-1}, i_{j-1}, or m_{j-1}, depending on whether the current state is a deletion, insertion, or match state. When the current state is a match state, the SMURF template replaces the transition$[u_{j-1}, m_j]$ term with a one. The β-strand component is only calculated when the particular match state m_j participates in a β-strand that is matched with a state m_k earlier in the sequence template. In fact, this component is the main difference between our MRF and an ordinary HMM. The only other difference is that we have modified the typical profile HMM to not allow insertion states along β-strands. So the transition probabilities above come directly from HMM training, except

$$\text{transition}[m_j|m_{j-1}] = 1 \text{ if } m_{j-1}, m_j \text{ in the same } \beta\text{-strand}$$

$$\text{transition}[m_j|d_{j-1}] = 1 \text{ if } m_j \text{ is in a } \beta\text{-strand}$$

We also can write down the probabilities of entering the *j*th deletion or insertion state; they are the same as for an ordinary HMM, except, again recall that we have modified the typical profile HMM not to allow insertion states along β-strands (as noted below):

$$\Pr[d_j|m_{j-1}] = \text{transition } [d_j, m_{j-1}]$$

where transition $[d_j, m_{j-1}] = 0$ if m_j is in a β-strand, and from the HMM transition probabilities otherwise:

$$\Pr[d_j|d_{j-1}] = \text{transition } [d_j, d_{j-1}]$$

$$\Pr[d_j|i_{j-1}] = 0$$

$$\Pr[i_j|m_j] = \text{transition } [i_j, m_j] \cdot \text{HMM}[i_j, r_k]$$

$$\Pr[i_j|i_j] = \text{transition } [i_j, i_j] \cdot \text{HMM}[i_j, r_k]$$

$$\Pr[i_j|d_j] = 0$$

Finally, the probability of the start state is calculated similar to the probability of assigning m_j to r_k, except the score is multiplied by a constant dependent on m_j. The probability of the end state is simply a constant dependent on m_j. Just like HMMER, we divide all probabilities by the probability of the "null model" when calculating the actual score. The null model is the probability of seeing the sequence by chance, based on the background residue frequency in proteins in general. Note that when actually calculating these scores, we instead calculate the logs, so that all products become summations. The score of the sequence is the maximum score obtained over all possible ways to parse the sequence onto the states of the MRF. It is converted to a *p* value by fitting a Gaussian to the scores of the nonpropeller containing sequences of length at least 150. For the double-propeller templates, formed by chaining together two SMURF single-propeller templates, it is difficult to calculate *p* values directly because the set of solved structures is biased toward shorter protein chains, and the YYY proteins are very long. Thus we calculate *p* values as follows: We add the means and standard deviations used to calculate the component single-propeller Gaussians to create a new Gaussian.

The maximum score of a sequence is computed by multidimensional dynamic programming. The dynamic programming takes place on the MRF described above— it has the same start states, end states, match states, insertion states, and deletion states as does HMMER HMMs, except some of the states are special β-strand states. Recall that the β-strand states consist of sets of contiguous states with no insertion or deletion states allowed between them (the β-strands), and furthermore, β-strands are paired with other β-strands. In particular, the pairwise probabilities for paired β-strands can only be calculated for the second paired β-strand, once it has been fixed what residue will be occupying the first position of the pair. Thus, each time we reach a state in the HMM that corresponds to the first residue of the first β-strand in a set of paired β-strands, we need to keep track of multiple cases, depending

on what sequence position is mapped to that state in the dynamic program. We keep track of this using a multidimensional array. For arbitrary gap lengths, this quickly becomes computationally infeasible, so a maximum gap length is defined (and for our purposes is set to the longest gap seen in the training alignment plus 20). When paired β-strands follow each other in sequence with no interleaving β-strands between them (which is usually the case for the β-propellers) the number of dimensions in the table for the portion of the dynamic program that parses the parts of the HMM between the two β-strands is directly proportional to the maximum gap length.

We now introduce some notation that will help describe the algorithm and analyze its complexity for interleaving pairs of β-strands. When we look, in order of sequence, at the states in the template MRF, mark the first of a pair of β-strands with consecutive numbers 1, 2, 3, and so on, and mark its corresponding paired β-strand with the same number. Thus, if we have *n* pairs of β-strands, we get a sequence that contains precisely two occurrences of each of the numbers from one to *n*. Now replace the first occurrence of each number with a left parenthesis and the second occurrence with a right parenthesis. Starting with the number zero, walk along the sequence, adding one for every left parenthesis, and subtracting one for every right parenthesis. The maximum size of this total at any intermediate point we call the *interleaving number* of this sequence. We call the last MRF state for the first of each pair of β-strands the "split" state and the first MRF state for the second of that pair of β-strands the "join" state.

At every split state, the number of dimensions of the dynamic program will be multiplied by the maximum gap length, because the dynamic program has to keep track of scores for each possible sequence position (up to maximum gap length) that could be mapped to that state. At the corresponding join state, the number of dimensions will be reduced by the maximum gap length, because the scoring function can calculate all the pairwise probabilities (and hence the score) of placing that residue into the join state, and then simply take the maximum of all ways to have placed its paired residue into the split state. Thus the number of elements in the multidimensional table is never more than sequence length times the maximum gap length raised to the interleaving number power. As noted above, when there are few or no interleaving β-strand pairs, as is the case with the strand topology of the β-propellers, this is very computationally feasible. On the other hand, the algorithm may become computationally infeasible for structures with multiple interleaving β-strands, such as the parallel β-helix structures.

HMMER implementation. Our program was tested against HMMER version 3.0a2 (22) with the "–seqZ 1" and "–seqE 10, 000" options, and otherwise all default settings [including the sequence entropy weighting (30) which is the default in this version of HMMER; we also tried HMMER without sequence entropy, and performance was worse]. The –seqZ 1 option makes returned E-values comparable between runs on different sized sequence databases, and the –seqE 10, 000 option causes HMMER to return results for all input sequences. Note that running HMMER is not exactly the same as running SMURF and weighting the pairwise score zero and the HMM score one, because SMURF needs to constrain the HMM to have contiguous match states (with no insertions or deletions) between adjacent residues in each individual β-strand to also compute a pairwise score.

ACKNOWLEDGMENTS. This work was partially supported by National Institutes of Health Grant 1R01GM080330-01A1 (to L.C.).

1. Mascher T, Helmann JD, Unden G (2006) Stimulus perception in bacterial signal-transducing histidine kinases. *Microbiol Mol Biol R*, 70:910–938.
2. Sonnenburg E, et al. (2006) A hybrid two-component system protein of a prominent human gut symbiont couples glycan sensing in vivo to carbohydrate metabolism. *Proc Natl Acad Sci USA*, 103:8834–8839.
3. Wexler H (2007) Bacteroides: The good, the bad, and the nitty-gritty. *Clin Microbiol Rev*, 20:593–621.
4. Bateman A, et al. (1994) The Pfam protein families database. *Nucleic Acids Res*, 32:D138–141.
5. Kinderman R, Snell JL (1980) *Markov Random Fields and Their Applications* (American Mathematical Society, Providence, RI), pp 24–33.
6. Olmea O, Rost B, Valencia A (1999) Effective use of sequence correlation and conservation in fold recognition. *J Mol Biol*, 293:1221–1239.
7. Steward RE, Thorton JM (2002) Prediction of strand pairing in antiparallel and parallel β-sheets using information theory. *Proteins Struct Funct Bioinf*, 48:178–191.
8. Cowen L, Bradley P, Menke M, King J, Berger B (2002) Predicting the beta-helix fold from protein sequence data. *J Comput Biol*, 9:261–276.
9. Rost B, Sander C (1993) Prediction of protein secondary structure at better than 70% accuracy. *J Mol Biol*, 232:584–599.
10. Jones D (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol*, 292:195–202.
11. Lesk A, LoConte L, Hubbard T (2001) Assesent of novel fold targets in CASP4: Predictions of three-dimensional structures and interresidue contacts. *Proteins Struct Funct Bioinf*, 45:98–118.
12. Bradley P, et al. (2003) Rosetta predictions in CASP5: Successes, failures and prospects for complete automation. *Proteins Struct Funct Bioinf*, 53:457–468.
13. Srinivasan R, Rose G (2002) Ab initio prediction of protein structure using LINUS. *Proteins Struct Funct Bioinf*, 47:489–495.
14. Rost B (1999) Twilight zone of protein sequence alignment. *Protein Eng*, 12:85–94.
15. Bradley P, Cowen L, Menke M, King J, Berger B (2001) Betawrap: Successful prediction of parallel β-helices from primary sequence reveals an association with many microbial pathogens. *Proc Natl Acad Sci USA*, 98:14819–14824.
16. Liu Y, Carbonell J, Gopalakrishnan V, Weigele P (2009) Conditional graphical models for protein structural motif recognition. *J Comput Biol*, 16:639–657.
17. Liu Y, Carbonell J, Weigele P, Gopalakrishnan V (2005) Segmentation conditional random fields (SCRFs): A new approach for protein fold recognition. *Research in Computational Molecular Biology*, Lecture Notes in Computer Science (Springer, Berlin), 3500, pp 408–422.
18. McDonnell A, et al. (2006) Fold recognition and accurate sequence-structure alignment of sequences directing beta-sheet proteins. *Proteins Struct Funct Bioinf*, 63:976–985.
19. Menke M, King J, Berger B, Cowen L (2005) Wrap and pack: A new paradign for beta structural motif recognition with application to recognizing beta trefoils. *J Comput Biol*, 12:777–795.
20. Waldispuhl J, O'Donnell C, Devadas S, Clote P, Berger B (2007) Modeling ensembles of transmembrane beta-barrel proteins. *Proteins Struct Funct Bioinf*, 71:1097–1112.
21. Fullop V, Jones D (1999) Beta propellers: Structural rigidity and functional diversity. *Curr Opin Struc Biol*, 9:715–721.
22. Eddy S (2008) A probabilistic model of local sequence alignment that simplifies statistical significance estimation. *PLoS Comput Biol*, 4:e1000069.
23. Suzek B, Huang H, McGarvey P, Mazumder R, Wu C (2009) Uniref: Comprehensive and non-redundant Uniprot reference clusters. *Bioinformatics*, 23:1282–1288.
24. Chaudhuri I, Söding J, Lupas AN (2008) Evolution of the beta-propeller fold. *Proteins*, 71:795–803.
25. Jawad Z, Paoli M (2002) Novel sequences propel familiar folds. *Structure*, 10:447–454.
26. Paoli M (2001) Protein folds propelled by diversity. *Prog Biophys Mol Biol*, 76:103–130.
27. Canutescu A, Shelenkov A, Dunbrack RD (2003) A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci*, 9:2001–2014.
28. Sayle R, Milner-White E (1995) RASMOL: Biomolecular graphics for all. *Trends Biochem Sci*, 20:374–376.
29. Menke M, Berger B, Cowen L (2008) Matt: Local flexibility aids protein multiple structure alignment. *PLoS Comput Biol*, 4:e10 doi:10.1371/journal.pcbi.0040010.
30. Karplus K, Barrett C, Hughey R (1998) Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, 14:846–856.
31. Xu J, Berger B (2006) Fast and accurate algorithms for protein side-chain packing. *J Assoc Comput Mach*, 53:533–557.