J Mol Model (2006) 12: 355–361
DOI 10.1007/s00894-005-0052-4

ORIGINAL PAPER

W. Michael Brown · Shawn Martin · Joseph P. Chabarek · Charlie Strauss · Jean-Loup Faulon

Prediction of β-strand packing interactions using the signature product

Received: 2 August 2005 / Accepted: 23 September 2005 / Published online: 7 December 2005

Abstract The prediction of β-sheet topology requires the consideration of long-range interactions between β-strands that are not necessarily consecutive in sequence. Since these interactions are difficult to simulate using *ab initio* methods, we propose a supplementary method able to assign β-sheet topology using only sequence information. We envision using the results of our method to reduce the three-dimensional search space of *ab initio* methods. Our method is based on the signature molecular descriptor, which has been used previously to predict protein–protein interactions successfully, and to develop quantitative structure–activity relationships for small organic drugs and peptide inhibitors. Here, we show how the signature descriptor can be used in a Support Vector Machine to predict whether or not two β-strands will pack adjacently within a protein. We then show how these predictions can be used to order β-strands within β-sheets. Using the entire PDB database with ten-fold cross-validation, we have achieved 74.0% accuracy in packing prediction and 75.6% accuracy in the prediction of edge strands. For the case of β-strand ordering, we are able to predict the correct ordering accurately for 51.3% of the β-sheets. Furthermore, using a simple confidence metric, we can determine those sheets for which accurate predictions can be obtained. For the top 25% highest confidence predictions, we are able to achieve 95.7% accuracy in β-strand ordering.

Keywords β-sheets · Secondary structure prediction · Signature descriptor · Support vector machine

W. M. Brown · S. Martin (✉) · J. P. Chabarek · J.-L. Faulon
Computational Biology, 9212, Sandia National Laboratories,
P.O. Box 5800, MS 310,
Albuquerque, NM 87185, USA
e-mail: smartin@sandia.gov
Tel.: +1-505-2843601

C. Strauss
Biosciences Division, Los Alamos National Laboratory,
Los Alamos, NM 87544, USA

Introduction

Despite large-scale efforts in the development of new methodologies and algorithms to predict protein structure accurately from amino-acid sequence, the problem remains unsolved. The fact that the amino-acid sequence of a protein dictates the acquisition and stabilization of its structure suggests that empirical rules can be derived relating sequence to structure, thereby helping to circumvent the enormous search space required by an atomistic molecular mechanics treatment. Indeed, the use of machine learning to generalize sequence–structure relationships has become a prominent approach in the development of algorithms for protein structure prediction [1], and continues to offer knowledge and rules with the potential for improvement in the accuracy of structure prediction. The sequence–structure relationships obtained using machine learning can be applied either as a method to reduce the search space of prediction algorithms or as a post-processing step to select between alternate conformations with small free-energy differences—an issue which has been attributed as a considerable difficulty in fold-prediction algorithms [2].

Perhaps the main focus of machine-learning within the structure-prediction problem has been aimed towards the development of algorithms for secondary-structure prediction. Classification of regions of protein sequence into α-helices, β-strands, and loops can now be achieved with over 75% accuracy [3–5]. Successful prediction of secondary-structure allows for an efficient hierarchical strategy for structure prediction in proteins with low homology to those of known structure. If we classify a protein structure as an arrangement of secondary-structure elements (SSEs) and the topology of the loops between them, structure prediction can be divided into the process of determining the structure of SSEs within a protein, followed by optimization of the SSE arrangement to form a domain or protein. Such strategies seem to have roots in nature, as compounding experimental evidence suggests that secondary-structure formation is an early event in protein folding [2].

In contrast to the prediction of α-helices and β-strands, which can be performed on consecutive sequences of amino acids with a propensity for adopting certain secondary structures, the prediction of β-sheet topology requires the consideration of long-range interactions between β-strands that are not necessarily consecutive in sequence. Unfortunately, these interactions are difficult to simulate using *ab initio* methods [6]. This difficulty makes it desirable to develop methods capable of assigning β-sheet topology from sequence information alone, reducing the space to be searched by structure-prediction algorithms [7]. As reviewed by Siepen et al. [8] several successful strategies for predicting β-sheet topology have been developed. These include algorithms for the prediction of the type and location of β-turns [9], the determination of strand register in β-sheets [10–12], and the prediction of edge strands within β-sheets [8, 13]. Here, we present a complementary method capable of predicting whether or not two β-strands will pack adjacently within a given β-sheet domain. Our method can use this information to predict β-strand ordering within a β-sheet or to predict β-sheet edge strands.

In previous work, we have used the signature molecular descriptor [14, 15] to encode molecular structure for use in regression and classification problems. We have also used the signature descriptor to predict protein–protein interactions, achieving 70–80% cross-validation accuracy on datasets containing known interactions for human, mouse, yeast, and *H. pylori* [16]. In the work presented here, we describe the application of this method to the problem of predicting β-strand packing interactions. The steps in the approach, illustrated in Fig. 1, require knowledge of the sequences of β-strands within a protein or domain (as obtained from secondary structure prediction, homology-based approaches, or simulation). Once these strands are obtained, they are encoded in a signature vector space, and all strand pairs are classified using an SVM with a customized "signature product" kernel. Due to the sparse nature of the encoding, we are able to process very large datasets. The classifier predicts which strands will pack adjacently within a protein. These predictions could be used to build β-sheet domains in order to reduce the search space of *ab initio* methods or, as a post-processing step to score potential folds generated by structure-prediction methods. When given a domain of β-strands, we show how a simple algorithm can be applied to predict the ordering of the β-strands to form a β-sheet. Of course, this ordering also allows us to predict (trivially) the two edge strands within the β-sheet. In addition to providing a prediction of β-strand ordering, the associated β-sheet score for the prediction gives us a measure of the confidence in that prediction—a key issue in any algorithm making predictions based on a training dataset.

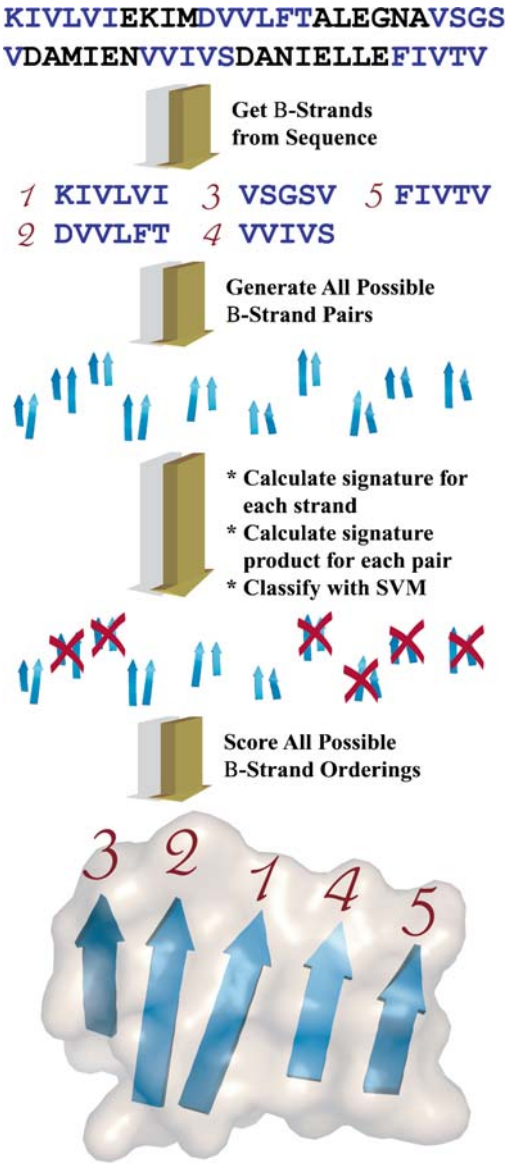

Fig. 1 Using signature products to predict the packing interactions within a β-sheet

Materials and methods

Signature molecular descriptor

The signature molecular descriptor provides a method for encoding two-dimensional molecular structure in a vector space [15]. Signature is based on the molecular graph of a molecule, where the vertices denote atoms in the molecule, and the edges correspond to the bonds between atoms. In this context, a molecule is characterized by a set of canonical subgraphs, each rooted on a different vertex with a predefined level of branching referred to as the height *h*. The branching of a vertex is an extended-degree sequence that describes the local neighborhood, up to a distance *h* away from the root. The signature for each atom is a canonical text representation of its corresponding subgraph.

Signature can be formulated as a function s:{molecular structure}→F defined by:

$$s(A) = \sum_i \sigma_i \mathbf{z}_i, \tag{1}$$

where A is a molecular structure, $\mathbf{z}_i$ is a basis vector in the signature space $F \cong \mathbb{R}^N$, and σ_i is the number of occurrences of $\mathbf{z}_i$ in A. In this case, as in the case of protein–protein interactions, we use amino-acid residues in place of atoms. This simplifies the molecular graph such that all except the end vertices have degree 2 and the signature space is described by subsequences of the β-strand's primary structure. A height 0 signature for a peptide consists of a count of the number of each of the amino-acid residue types present in the strand. A height 1 signature counts each of the possible tri-mers present in the peptide. A height 2 signature counts each of the possible five-mers present in the peptide, and so forth. The choice of the signature height depends on the specific problem. In our experience the best heights are usually 0, 1, or 2. For the β-strand problem, we found height 1 to provide the best test set accuracy (results not shown), and therefore consider only height 1 signatures in this paper. An example illustrating the height 1 signature for the peptide sequence LVMTTMK is shown in Fig. 2.

Signature product

The signature descriptor as described above is designed to encode only a single amino-acid sequence for a given data point. Here, however, we are interested in pairs of amino-acid sequences. Fortunately, there exists an extension of signature to sequence pairs *via* the signature product [16]. The signature product encodes a pair of amino-acid sequences as the tensor product between their individual signatures. The product signature for two β-strands A and B is given by:

$$\Gamma(A,B) = s(A) \otimes s(B) + s(B) \otimes s(A) \tag{2}$$

where the tensor product between two vectors $\boldsymbol{a}=(a_1,\ldots,a_n)^{\mathrm{T}} \in R^n$ and $\mathbf{b}=(b_1,\ldots,b_n)^{\mathrm{T}} \in R^m$ is given by $\mathbf{a}\otimes\mathbf{b}=(a_1b_1,a_1b_2,\ldots, a_nb_m)^T \in R^{nm}$. We use the sum on the right-hand side of Eq. 2 to enforce symmetry such that $\Gamma(A,B)=\Gamma(B,A)$.

Fig. 2 Height 1 signature for the peptide LVMTTMK. All trimers not shown have a coefficient of 0

In order to evaluate the signature product efficiently, a customized SVM kernel, k, can be used to compute the dot product between two β-strand pairs (A,B) and (C,D):

$$k((A,B),(C,D)) = \Gamma(A,B) \cdot \Gamma(C,D) \tag{3}$$

Furthermore, this kernel can be evaluated without the use of tensor products: [16]

$$k((A,B),(C,D)) = 2\begin{pmatrix} (s(A)\cdot s(C))(s(B)\cdot s(D)) + \\ (s(A)\cdot s(D))(s(B)\cdot s(C)) \end{pmatrix} \tag{4}$$

This identity allows us to evaluate the kernel while reducing the computational complexity introduced by using Eq. 2. For a detailed discussion on the signature product, see Martin et al. [16].

Support vector classification

Support-vector classification was performed using SVM$^{\mathrm{light}}$ [17] with kernel modifications as described above. A patch for SVM$^{\mathrm{light}}$, along with the necessary software for calculating signature products for protein sequences can be downloaded from http://www.cs.sandia.gov/~smartin/. An SVM was trained with cross-validation on non-homologous proteins (see below), using the linear signature-product kernel described above and default regularization.

Dataset

Training and testing were performed using β-strands extracted from a 2004 release of the RCSB Protein Data Bank (PDB) [18]. Any protein with over 95% homology to another protein in the dataset was removed, giving 6,682 proteins. For cross-validation, a random ordering of the proteins was divided into ten test sets, each consisting of approximately 10% of the proteins (668). All β-strand sequences, as assigned within the PDB records, were extracted and β-strand pairs were generated for every possible combination of β-strands within any given β-sheet. From these, all duplicate pairs, generated from multiple subunits, were removed. In addition, pairs containing less than four residues and greater than 100 residues were removed. Finally, a random selection of non-adjacent strands was removed to balance the number of adjacent and non-adjacent strands for unbiased training. The resulting set was composed of 27,196 adjacent strands and 27,196 non-adjacent strands to be used for training, with each cross-validation fold consisting of approximately 90% of these pairs.

For the non-homologous protein dataset, all β-sheets were isolated for validation of strand-ordering accuracy. Any β-sheet containing less than three strands, strands with less than four residues, strands with greater than 100 residues, or strands with unnatural amino-acid residues, was

removed. Duplicate β-sheets were also removed. It was verified that no strands in the test sets were present in the training sets. The cross-validation accuracy of β-strand pairing prediction was performed on all possible pairs in each sheet of the test sets. β-strand ordering based on the test-set predictions was also performed using the same 10-fold cross-validation.

β-strand ordering and edge strand prediction

We used a simple algorithm to predict the order of β-strands within a β-sheet. First, all possible strand pairs within the sheet were classified using the SVM model. The sign of the SVM prediction provided the class (adjacent or non-adjacent) and the magnitude gave the "strength" of the prediction. Next, we enumerated all possible orderings of the β-strands within the β-sheet. For each of these orderings, we arranged the SVM predictions into a packing-interaction matrix. This matrix is a symmetric matrix with one row (and one column) for each β-strand, where the row (and column) order is given by the proposed β-strand ordering. From the packing interaction matrix we derived two scores: a packing likelihood score, which was the average of the super-diagonal elements of the matrix (the elements directly above the diagonal); and a non-packing likelihood score, which was the average of the upper diagonal elements, not including the diagonal or the super-diagonal. The total score for a β-sheet is given by the packing-likelihood minus the non-packing likelihood of opposite sign. As an example, consider the β-sheet ordering of "3, 1, 2." This ordering would be probable if the elements on the super-diagonal ("3, 1" and "1, 2") of the packing interaction matrix were highly positive, and the upper diagonal element ("3, 2") of the matrix was highly negative. Edge strands are "3" and "2."

Our score can also be used as a confidence metric. This is true because the output of an SVM for a given strand pair can be interpreted as a confidence [19]. In particular, the output of a SVM is the distance of the input from a separating hyperplane. Therefore, classifications of β-strand pairs that have very small distances from the hyperplane will be less accurate than those with very large distances. β-strand pairs near the hyperplane have similarity to both pairing and non-pairing strands. Because the β-sheet score presented here is a summation of the mean magnitudes of the distances of pairing-strands from the hyperplane (packing likelihood) and distances of non-pairing strands from the hyperplane (non-packing likelihood), it not only represents a score for a given β-sheet ordering, but also a confidence in that score.

Results

Prediction of β-strand packing

In order to assess the accuracy of our method for predicting which β-strand pairs pack adjacently within a protein, we trained our SVM on ten folds using approximately 24,400 adjacent β-strands and 24,400 non-adjacent β-strands for training for each fold (note that these counts represent approximately 90% of the total dataset). The calculations were done in two steps. We first pre-computed the signature kernels (not the signature-product kernels) for use by the SVM. We next trained our SVM using the signature-product kernel. The resulting models misclassified an average of 26.6% of the training set. Classification of the test sets (with β-strand pairs extracted from 10% of the PDB in each case) resulted in a ten-fold cross-validation accuracy of 74.0%.

Prediction of β-strand ordering and edge strands

We next tested the ability of our method to predict the ordering of β-strands within a β-sheet. We benchmarked our method by using the strand orderings for all the β-sheets in the PDB that met our criteria (see Materials and methods). Choosing the correct β-sheet as the ordering that resulted in the highest score, we achieved an overall 10-fold cross-validation accuracy of only 49.3%. The accuracy for three-stranded sheets was 63.4%, for four-stranded sheets 56.58%, and for nine-stranded sheets 11.11%. The decrease in accuracy with sheet size is simply due to the increase in the number of classifications required to compute a score and the increase in the number of possible orderings for any given sheet. For example, an n-stranded sheet requires $1/2n(n-1)$ strand-pairing classifications and has $n!/2$ possible orderings. For a sheet with nine strands, 36 strand-pair classifications are required to calculate the

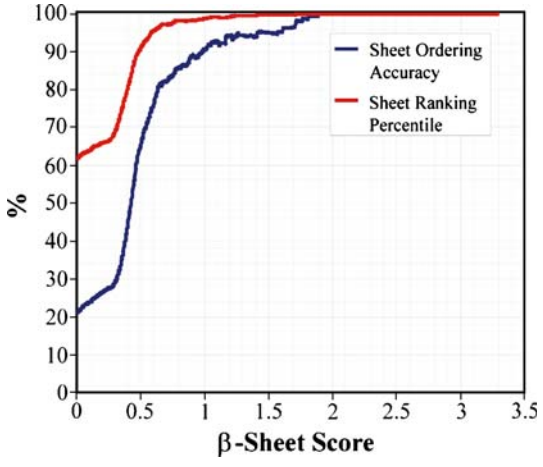

Fig. 3 Moving average plot of the ten-fold cross-validation β-sheet ordering accuracy and β-sheet rank percentile for the non-homologous PDB dataset as a function of the β-Sheet Score. The window for the moving average is 0.5. As the magnitude of the β-sheet ordering score increases, so does the confidence that the ordering is correct. For example, if the best score for all possible orderings of a β-sheet is 1.5, there is an expected 95% chance that this is the correct ordering and on average, 99.7% of incorrect orderings will be scored lower (based on the ten-fold cross-validation accuracy of β-sheets with scores between 1.25 and 1.75)

Table 1 Ten-fold cross-validation β-sheet ordering accuracy, sheet rank percentile, and edge strand accuracy for the non-homologous PDB dataset divided by the β-sheet score quartiles

Score	Percent of dataset	Ordering accuracy (%)	Sheet rank percentile (‰)	Edge strand accuracy (%)
≥0.8	25	95.65	99.53	98.23
0.21–0.8	25	60.73	89.09	81.78
0.11–0.21	25	20.84	66.14	62.59
<0.11	25	19.92	55.23	59.70
All	100	49.30	77.36	75.55

For the 25% of the dataset which was scored with the highest confidence (Score >0.8), the sheet ordering accuracy is 95.65% for β-sheets of all sizes. On average, 99.53% of β-sheets were scored below the correct ordering and edge strands were predicted with 98.23% accuracy

score and the correct sheet must be selected from 181,440 possible orderings.

An overall accuracy of 49.3% is impressive considering the difficulty of the problem and the fact that the baseline accuracy using randomly generated strands was 13.6% (note that this calculation takes into account the high percentage of 3- and 4-stranded β-sheets). However, the result is sub-optimal in terms of the end-user objective of generating orderings for protein-structure predictions. The problem, which plagues all machine learning models, centers around the critical question: "How do we know for which test cases the prediction accuracy will be high and for which test cases the prediction accuracy will be low?" This question has recently been addressed in the cheminformatics arena with the development of discriminators for prediction accuracy in quantitative–structure activity relationships (QSARs) [20]. In this work the authors found that the highest prediction accuracy was obtained when the test molecule in question had a large number of similar molecules in the training set. Since we are using SVM classification instead of general QSAR regression, we can use an approach inherent to the SVM model itself. Due to the fact that the β-sheet score is a measure of the mean distance of the pairing and non-pairing strands from the hyperplane, we can use this score not only to predict the correct ordering, but also as a measure of confidence in that prediction (described in the Materials and methods section).

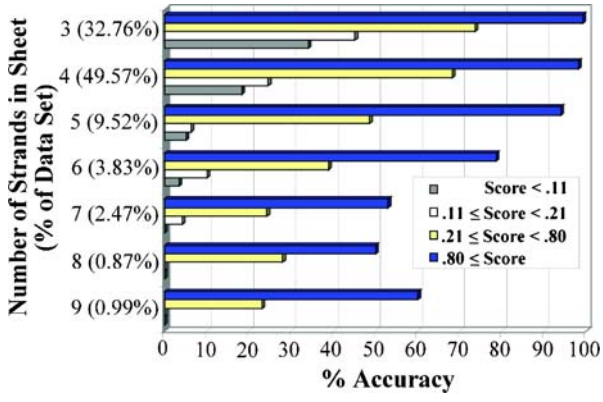

Fig. 4 Ten-fold cross-validation prediction accuracy for β-strand ordering as a function of the number of strands within a sheet for the non-homologous PDB dataset divided by the β-sheet score quartiles. The numbers in parenthesis represent the percentage of β-sheets within the dataset containing that number of strands

It turns out that this confidence metric correlates surprisingly well with prediction accuracy, as shown in Fig. 3 and Table 1. In Table 1, we divide the dataset into four equally sized subsets based on the β-sheet score quartiles (0.11 for the lower quartile, 0.21 for the median quartile, and 0.8 for the upper quartile), and recalculate the accuracies. For the 25% of the predictions with the highest confidence, a 95.7% ordering accuracy was achieved, while for the bottom 25%, the accuracy was 19.92%. A breakdown of the accuracies by the size of the β-sheet is given in Fig. 4.

The results show that for about 1 in 4 of the β-sheets encountered in the PDB, the β-sheet score is sufficient to have high confidence in the predicted ordering. What if, however, we are interested in a β-sheet with a lower confidence score? Under these circumstances, it may not be appropriate to select only one β-sheet ordering as correct, but rather remove those β-sheet orderings that are highly unlikely. For these cases, a β-sheet ranking percentile is appropriate. We calculate the ranking percentile as the average percentage of β-sheet orderings that score below the correct one. Using this approach, we can remove on average from 55% to 90% of the alternate orderings (Table 1, Fig. 3).

For the prediction of edge strands, an overall accuracy of 75.6% is obtained. As with the β-sheet ordering accuracy, the confidence correlates with the β-sheet score. For the top 25% of the database, the edge strand prediction accuracy is 98.2% and for the bottom 25% it is 59.7% (Table 1). As is to be expected, there is a decrease in prediction accuracy with an increase in the number β-strands within a given sheet (data not shown).

Discussion

The signature molecular descriptor has been applied successfully to the elucidation of quantitative structure activity relationships for log P, HIV organic drug IC50s, and peptide inhibitor IC50s, [14, 15] as well as to classification of protein–protein interactions [16]. Here we have shown how the descriptor can be used to predict β-sheet topology based on sequence information alone. Using the signature product SVM, we can predict whether or not two β-strands will pack adjacently within a protein. We used the entire PDB database to validate our method and achieved an overall accuracy of 74.0%. When given the strands within a

β-sheet, the model can predict the ordering with an overall accuracy of 49.3%. However, using a simple prediction confidence metric, we can determine *a priori* when the accuracy of a prediction should be high enough to trust as a correct ordering. For test cases where the confidence is low or where the number of β-strands in the sheet is high, the model is not sufficient for predicting β-strand ordering as a starting point for *ab initio* protein structure prediction methods. Rather, it can be used to throw out potential folds that are predicted to be highly unlikely. On average, 77.36% of the possible β-strand orderings were predicted with a lower score than the correct one.

Our model is not directly related to existing methods, although we can offer some qualitative comparisons. The most closely related method in terms of ultimate goal is Golem, as described in King et al. [13]. Golem uses inductive learning to discover topological rules related to the packing of β-sheets, as well as to determine which β-strands are at the edges of a given sheet. However, Golem operates on a small scale relative to our approach (only a few structures were considered), and uses an entirely different computational approach. The most closely related work in terms of method is that described by Siepen et al. [8] where SVMs and decision trees were used to predict when a β-strand is an edge strand of a β-sheet. In their work, a large dataset was used and an accuracy of 78% was reported for predicting edge strands. Our method for edge-strand prediction gave an overall accuracy of 75.6%; however, it cannot be compared to the results of Siepen et al., because their method can predict edge strands from any set of β-strands, while our method can predict edge strands only within a given β-sheet. Finally, other existing methods compute rather different topological properties, such as the methods of Hutchinson et al. [10] Zaremba and Gregoret [12] and Steward and Thornton [11] for predicting strand register and orientation (parallel or anti-parallel), and the method of Przybylski and Rost [9] for locating β-turns.

In ordering β-strands within sheets, our model is able to predict the edge strands of the sheet with an overall accuracy of 75.6%. This is relevant for two reasons. First, from a biological standpoint, recognition between exposed edge strands of β-sheets is an important mode of protein–protein interaction; edge strands are implicated in the aggregation of engineered β-sheet proteins; and the interaction between the edges of β-sheets has been linked intimately to the aggregation of proteins into pathogenic cross-β fibril structure, which is associated with a number of disorders (reviewed by Siepen et al. [8]). Therefore, knowledge of the edge strands within a protein provides useful information beyond the realm of predicting sheet topology. Second, from a computational point of view, 24.4% of our β-strand ordering predictions were incorrect because they did not position the edge strands correctly. Therefore models able to predict edge strands correctly could potentially provide a significant improvement in our ordering accuracy.

Several unique characteristics of edge strands have been observed, possibly a result of evolution to protect edge strands from interactions with other β-strands [21]. These include the presence of charged residues, proline residues, β-bulges, and a lack of alternating periodicity of hydrophobic and polar residues commonly seen in β-strands. Siepen et al. [8] used these unique characteristics of edge strands, along with the observation that shorter strands tend to be edge strands, in order to develop a model for edge strand prediction that had an accuracy of approximately 78%. Because these characteristics are not likely to be well represented in our model, which included no explicit training for edge-strand prediction, incorporation of similar descriptors for the purpose of edge-strand prediction may lead to further improvement in the accuracy of our method for β-strand ordering. Finally, in addition to improving our model, we intend to test our approach by predicting other topological properties of proteins, including β-strand orientation, prediction of edge strands without knowledge of β-sheet domains, prediction of α-helix packing interactions (which is especially relevant for structure prediction of integral membrane proteins [22]), and possibly, prediction of contact points.

Acknowledgements This work was funded by the U.S. Department of Energy's Genomics: GTL program (http://www.doegenomestolife.org) under project, "Carbon Sequestration in Synechococcus Sp.: From Molecular Machines to Hierarchical Modeling" (http://www.genomes-to-life.org). Sandia is a multiprogram laboratory operated by Sandia Corporation, a LockheedMartin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

References

1. Bohm G (1996) Biophys Chemist 59:1–32
2. Honig B (1999) J Mol Biol 293:283–293
3. Jones DT (1999) J Mol Biol 292:195–202
4. Lin K, Simossis VA, Taylor WR, Heringa J (2005) Bioinformatics 21:152–159
5. Rost B (2001) J Struct Biol 134:204–218
6. Orengo CA, Bray JE, Hubbard T, LoConte L, Sillitoe I (1999) Proteins (Suppl 3):149–170
7. Kolinski A, Betancourt MR, Kihara D, Rotkiewicz P, Skolnick J (2001) Proteins 44:133–149
8. Siepen JA, Radford SE, Westhead DR (2003) Protein Sci 12: 2348–2359
9. Przybylski D, Rost B (2002) Proteins 46:197–205
10. Hutchinson EG, Sessions RB, Thornton JM, Woolfson DN (1998) Protein Sci 7:2287–2300
11. Steward RE, Thornton JM (2002) Proteins 48:178–191
12. Zaremba SM, Gregoret LM (1999) J Mol Biol 291:463–479
13. King RD, Clark DA, Shirazi J, Sternberg MJ (1994) Protein Eng 7:1295–1303
14. Churchwell CJ, Rintoul MD, Martin S, Visco Jr DP, Kotu A, Larson RS, Sillerud LO, Brown DC, Faulon JL (2004) J Mol Graph Model 22:263–273
15. Faulon JL, Visco Jr DP, Pophale RS (2003) J Chem Inf Comput Sci 43:707–720
16. Martin S, Roe D, Faulon JL (2005) Bioinformatics 21:218–226

17. Joachims T (1999) In: Scholkopf B, Burges CJC, Smola AJ (eds) Advances in Kernel Methods-Support Vector Learning, pp 169–184
18. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) Nucleic Acids Res 28:235–242
19. Dumais ST (1998) IEEE Intelligent Systems Magazine 13: 21–23
20. Sheridan RP, Feuston BP, Maiorov VN, Kearsley SK (2004) J Chem Inf Comput Sci 44:1912–1928
21. Richardson JS, Richardson DC (2002) Proc Natl Acad Sci USA 99:2754–2759
22. Brown WM, Faulon JL, Sale K (2005) Comput Biol Chem 29: 143–150