β Edge strands in protein structure prediction and aggregation

JENNIFER A. SIEPEN, SHEENA E. RADFORD, AND DAVID R. WESTHEAD
School of Biochemistry and Molecular Biology, University of Leeds, Leeds LS2 9JT, UK
(RECEIVED May 30, 2003; FINAL REVISION July 9, 2003; ACCEPTED July 11, 2003)

Abstract

It is well established that recognition between exposed edges of β-sheets is an important mode of protein–protein interaction and can have pathological consequences; for instance, it has been linked to the aggregation of proteins into a fibrillar structure, which is associated with a number of predominantly neurodegenerative disorders. A number of protective mechanisms have evolved in the edge strands of β-sheets, preventing the aggregation and insolubility of most natural β-sheet proteins. Such mechanisms are unfavorable in the interior of a β-sheet. The problem of distinguishing edge strands from central strands based on sequence information alone is important in predicting residues and mutations likely to be involved in aggregation, and is also a first step in predicting folding topology. Here we report support vector machine (SVM) and decision tree methods developed to classify edge strands from central strands in a representative set of protein domains. Interestingly, rules generated by the decision tree method are in close agreement with our knowledge of protein structure and are potentially useful in a number of different biological applications. When trained on strands from proteins of known structure, using structure-based (Dictionary of Secondary Structure in Proteins) strand assignments, both methods achieved mean cross-validated, prediction accuracies of ~78%. These accuracies were reduced when strand assignments from secondary structure prediction were used. Further investigation of this effect revealed that it could be explained by a significant reduction in the accuracy of standard secondary structure prediction methods for edge strands, in comparison with central strands.

Keywords: Decision tree; support vector machine; secondary structure prediction; machine learning

β-sheets, comprising two or more parallel or antiparallel β-strands connected by interstrand hydrogen bonds, are a well-known feature of many protein structures. Two distinct tertiary contexts of a β-strand may be defined: "central strands, bordered on both sides by other β-strands, and edge strands, bordered on only one side by another β-strand" (Minor and Kim 1994). It is well established that recognition between exposed edge strands of β-sheets is an important mode of protein–protein interaction; for example, the amphiphilic dimer formed by Defensin HNP-3 (Fig. 1) at which edge strands from the three-stranded sheet in each monomer connect by four hydrogen bonds to form a six-stranded β-sheet (taken from the Protein Data Bank; Hill et al. 1991; Berman et al. 2000).

This effect is also implicated in the aggregation of designed β-sheet proteins. For example, Wang and Hecht (2002) designed a combinatorial library of de novo six- and eight-stranded β-sheet proteins. The designed proteins, consisting of identical β-strands designed to match the natural structural periodicity of amphiphilic β-strands, favored intermolecular oligomerization, and the β-sheet proteins formed amyloid-like fibrils. Specific redesign of this library of β-sheet proteins, by changing the binary pattern of the first and/or last β-strands of several sequences to include a

Reprint requests to: David R. Westhead, School of Biochemistry and Molecular Biology, University of Leeds, Leeds, LS2 9JT, UK; e-mail: westhead@bmb.leeds.ac.uk; fax: 44-113-343-3167.

Abbreviations: β2M, β-2-microglobulin;, DSSP, Dictionary of Secondary Structure in Proteins; HSSP, database of homology-derived secondary structure of proteins; MCC, Matthews correlation coefficient; PQS, protein quaternary structure; SVM, support vector machine; TTR, transthyretin.

Article and publication are at http://www.proteinscience.org/cgi/doi/10.1110/ps.03234503.

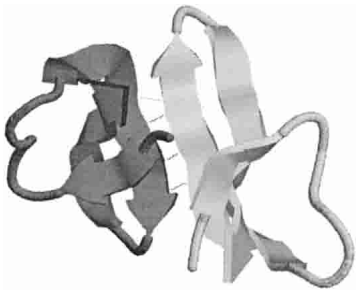

Figure 1. The Defensin HNP-3 dimer. RASMOL cartoon (Sayle and Milner-White 1995) representation of the Defensin HNP-3 dimer, 1DFN. The two monomers are in close contact, with four hydrogen bonds (dotted lines) between the edge strands of the two monomers at the dimer interface (Hill et al. 1991).

charged lysine residue, prevented oligomerization and favored monomeric β-sheet proteins (Wang and Hecht 2002). Similarly, a computer algorithm has been used to search for short sequences with a high propensity to form homopolymeric β-sheets (Lopez de la Paz et al. 2002). Sequences predicted to be favorable for such interactions (predominantly based on the overall charge of the sequence) were found to form β-sheets experimentally, whereas the introduction of specific point mutations into the sequences inhibited polymerization (Lopez de la Paz et al. 2002).

Despite this, there are relatively few cases of aggregation or insolubility in natural β-sheet proteins, and recently, Richardson and Richardson (2002) described natural mechanisms that the edge strands of β-sheets have evolved to protect edge strands from interactions with other β-strands. These negative design mechanisms include charged residues, proline residues, and/or β bulges that are common in the edge strands of β-sheet proteins. These protective features are unfavorable in the interior strands of the β-sheet and, as a result, tend to lower the β-sheet propensity of the edge strand (Richardson and Richardson 2002). This and other studies (see Sternberg and Thornton 1977; King et al. 1994; Minor and Kim 1994) demonstrate characteristic differences, and different β-sheet propensities, between strands located toward the edge of the β-sheet and those occupying a more central location. Edge strands tend to lack the classic patterns often associated with β-strands, such as the alternating periodicity of hydrophobic and polar residues; they tend to be of less hydrophobic nature and to contain more charged residues (Sternberg and Thornton 1977; King et al. 1994; Minor and Kim 1994; Richardson and Richardson 2002; Wang and Hecht 2002).

The interaction between the edges of β-sheets has been intimately linked to the aggregation of proteins into pathogenic cross-β fibril structure, associated with a number of disorders. For example, the tetrameric thyroid hormone-transporter protein, transthyretin (TTR), associated with amyloid fibril deposition in familial amyloidotic polyneuropathy, forms monomeric and oligomeric intermediates at low pH, which can self-assemble into a fibril structure. Serag and coworkers (2001, 2002) have identified specific subunit interactions in the aggregated form of TTR. A head-to-head arrangement of subunits has an interface similar to that found in the native soluble form of TTR, in which the edge strands from the two monomers are hydrogen-bonded across the subunit interface (Serag et al. 2001). A second tail-to-tail arrangement of subunits follows major conformational changes, which displace the protective edge strand of the β-sheet in the native TTR structure, exposing the penultimate strand that forms the second subunit interface (Serag et al. 2002). Interestingly, the distribution of disease-associated mutational variants in TTR peaks in the edge strand observed to undergo this conformational change (Serpell et al. 1996). In addition to TTR, the role of edge strands in the fibrillogenesis of other proteins has been studied. For example, the dissociation of β2-microglobulin (β2M) from the heavy chain of the class I HLA complex is a critical first step in the formation of amyloid fibrils implicated in dialysis-related amyloidosis. The monomeric crystal structure of β2M reveals structural changes, relative to the HLA-bound form, that restructure two short edge strands connected by a β bulge into a single longer strand at one edge of the β-sheet, resulting in a surface potentially prone to aggregation (Trinh et al. 2002). The other edge strands of β2M, formed by the N and C termini, have been found to be weakly protected from hydrogen exchange in β2M amyloid fibrils, indicating that they are unstructured in the fibril (Hoshino et al. 2002). Furthermore, topological investigation of β2M fibrils by Monti et al. (2002) revealed that proteolytic processing after the formation of fibrils by β2M leads to an N-terminal truncated form of the protein.

These observations indicate that knowledge of the location of potential edge strands would be extremely useful. Traditional methods predict the secondary structure of proteins into three main classes (helix [H], extended [sheet, E], and coil [C]) with an accuracy of ~75% (Przybylski and Rost 2002). Prediction of more detailed local structure in proteins has also been investigated. For example, as an extension to a three-state secondary structure prediction, neural networks have been used to predict the location and type of β-turns in protein sequences (Shepherd et al. 1999). In addition, Bystroff and Baker (1998) predicted local structure in proteins by using a library (I-Sites) of sequence–structure motifs, and more recently, Pollastri et al. (2002) extended the classic three state classification into eight classes. Despite this, there are currently no commonly used prediction methods available that specifically predict the edge strands of β-sheets.

Different machine learning approaches have been used to predict the arrangement of the secondary structure elements within the protein, often after the assignment of secondary structure by secondary structure prediction methods. An in-

ductive logic programming approach, GOLEM, was used (King et al. 1994) to discover topological rules in the packing of β-sheets in α/β domain proteins; this included the generation of rules for the determination of whether or a not a strand was at the edge of a sheet. Although useful rules were generated, their accuracy in the prediction of unknown proteins was not evaluated, and the rules were only applicable within a relatively limited set of α/β protein structures (King et al. 1994). Related work has considered the determination of strand register in β-sheets, including context-dependent amino acid pairings (Zaremba and Gregoret 1999) and strand pairings in parallel and antiparallel β-sheets (Hutchinson et al. 1998, Steward and Thornton 2002), with some degree of success, but none of these have considered edge strand prediction.

The observations above indicate that it might be possible to predict whether a strand occupies a central or edge location in a β-sheet. Such a prediction method would be valuable in predicting mutations likely to influence aggregation. More generally, it would contribute to the prediction of the folding topology of the sheet and, thus, serve as a first step in tertiary structure prediction. Further, it could be a valuable aid in the evaluation of tertiary structure prediction by fold recognition, which is improved by the incorporation of secondary structure predictions (see Kelley et al. 2000), or ab initio methods (see Simons et al. 1997). Here we describe the application of two machine learning methods, support vector machines (SVMs) and decision trees, to this important prediction problem. SVMs are based on statistical learning theory developed by Vapnik et al. (1998), and in addition to their ability to generalize well, they have been shown to outperform other learning methods in a range of applications (Cristianini and Shawe-Taylor 2000). Decision trees, on the other hand, are able to produce human readable rules, which may be used to provide some meaningful explanations for how the data is classified.

Results

The machine learning methods known as SVMs and decision trees were used to distinguish edge strands from central strands based only on information derived from the amino acid sequences of strands concerned. A set (see Materials and Methods) comprising 564 nonhomologous proteins of known structure was used to provide training and cross-validation data. After optimization of the parameters and kernel function from the strand data with the SVM method (SVMTorch; Collobert and Bengio 2001; see Materials and Methods), a Gaussian kernel with 6 SD was chosen and used in all the SVM predictions reported below. The default parameters for the decision tree method (C4.5; Quinlan 1993) were found to be approximately optimal and used throughout.

To evaluate the performance of the machine learning methods, the technique of cross-validation was used. This is a procedure in which a data set containing instances of known class (strands known to be either edge or central in this case) is randomly split into two parts, a training set and test set. The term *n*-fold cross-validation is used when the data set (in randomized order) is split into *n* equal parts, each part being used in turn as test set with the remaining $n-$ parts as training set. In this case, the cross-validation experiments can be used to assess the variability (precision) of the estimated measures of performance, and in the work below, results are reported as mean ± SE.

When considering the performance of these prediction methods, it is important to note that a naive method predicting each class (edge, central) randomly with equal probability has an expected accuracy of 50% correct predictions on any test data set. Given information about the composition of the data set, it is straightforward to show that the best use a naive method could make of this would be to predict all instances to come from the dominant class (edge in this case), resulting in an accuracy equal to the percentage of instances of the dominant class in the data (52.9% in this case). To be useful, a method needs to exceed these naive methods significantly in the overall percentage of correct predictions. As well as the overall percentage of correct predictions, several other measures of performance were used in the work below. For each class of strand (edge or central), the *sensitivity* of a method is the percentage of the test set strands actually in that class that were correctly predicted to be in that class. On the other hand, the *specificity* is the percentage of the test set strands predicted to be in a given class that actually are in that class. Values <100% in *sensitivity* and *specificity* for the prediction of a particular class reflect the occurrence of false-negative and false-positive errors, respectively, for that class. In addition, the Matthews correlation coefficient (MCC; Matthews 1975; see Materials and Methods) was used as a robust, single-valued measure of performance, accounting for both false-positive and false-negative prediction errors for each class.

Prediction based on strand assignments from structure

Using secondary structure elements defined based on structural information from the Dictionary of Secondary Structure in Proteins (DSSP; Kabsch and Sander 1983) allowed us to investigate the performance of the prediction method independent of any confounding effect of inaccuracy in secondary structure prediction. The edge and central strands of each protein were determined by using hydrogen bonding information from DSSP (Kabsch and Sander 1983; see Materials and Methods). From the nonhomologous data set described above, this yielded a data set comprising the amino acid sequences of 3359 edge strands and 2995 central strands.

The literature cited above, in particular the work of Richardson and Richardson (2002), identified a number of features seen frequently in edge strands that are unfavorable in the interior of the β-sheet. These features were exploited in this method to enable the distinction of edge strands from central strands. The attributes investigated as inputs for the machine learning methods were strand length, hydrophobicity, hydrophobic moment, periodicity of polarity, proportion of charged residues, and the propensity of the strand to contain a β bulge (described in Materials and Methods). A preliminary investigation of the predictive power of these attributes used a ninefold cross-validation of the decision tree method. Attributes were added sequentially, as shown in Figure 2. It is clear from the figure that each added attribute improved the prediction accuracy, and the greatest improvement was observed when the proportion of charged residues in the strand and the periodicity of polarity (see Materials and Methods) were included. Based on these observations, all six attributes were used as input for the machine learning methods described below.

For the full assessment of performance, an 18-fold cross-validation technique was used for both the SVM and decision tree methods. The results are given in Table 1. With the SVM, 78.0 ± 0.4% (MCC, 0.59) of strands were classified correctly as edge or central strands by the SVM method based on the six attributes discussed above. A similar overall prediction accuracy of 78.0 ± 0.5% (MCC, 0.57) was achieved by the decision tree method. The sensitivity and specificity of both methods are high (>70% in all cases, and often >80%). The results of both machine learning methods are significantly better than the expected results for either of the naive methods described above (the difference in overall accuracy is significant at the 99% level in a *t* test).

Table 1. *The prediction accuracy of edge and central strands*

	Actual Strand Class		
Predicted strand class	Central strands	Edge strands	
SVM method			
Central strands	144 ± 2.4	21 ± 1.0	87 ± 0.6%
Edge strands	55 ± 1.2	131 ± 1.8	70 ± 0.5%
	72 ± 0.7%	86 ± 0.6%	78 ± 0.4%[b]
	Sensitivity		Specificity[a]
Decision tree method			
Central strands	137 ± 2.1	49 ± 1.4	74 ± 0.6%
Edge strands	28 ± 1.5	137 ± 2.5	83 ± 0.8%[a]
	83 ± 0.9%	74 ± 0.8%	78 ± 0.5%[b]
	Sensitivity[c]		

[a] The proportion of strands predicted to be central/edge strands that actually are central/edge strands.
[b] The overall percentage of predictions that were correct.
[c] The proportion of the central/edge strands predicted to be central/edge strands.

Multiple sequence alignments provide useful information about the most conserved regions of a protein sequence likely to be structurally important and/or buried within the protein core, and consequently, more information may be attained from an alignment rather than a single sequence. The incorporation of multiple sequence alignments into secondary structure prediction methods, promoting regular α/β structures in areas of high sequence conservation and penalizing the prediction of α/β structures in sections of low sequence conservation, leads to a significantly better prediction accuracy (see Zvelebil 1987). The Database of Homology-Derived Secondary Structure of Proteins (HSSP; Sander and Schneider 1991) contains sequence alignments

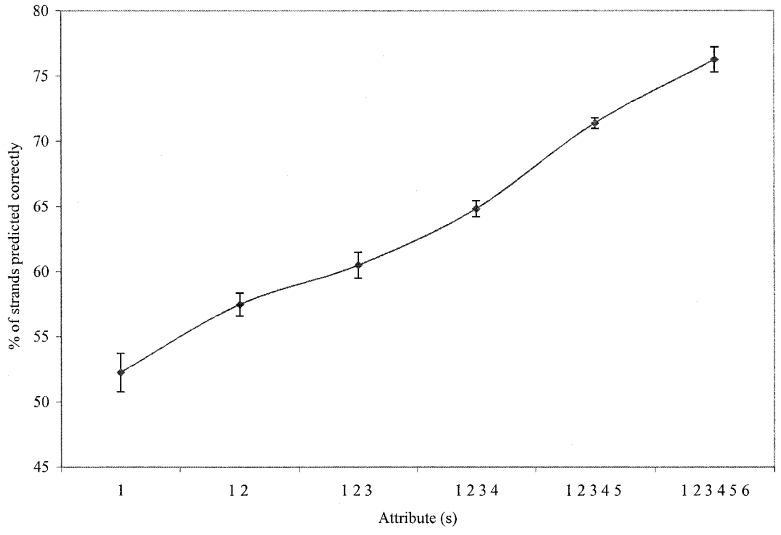

Figure 2. Effect of attribute selection. The effect of each of the added attributes on the percentage of the edge and central strands correctly classified based on a ninefold cross-validation technique. The attributes are (1) bulge score, (2) charge score, (3) hydrophobicity, (4) hydrophobic moment, (5) pattern of polarity, and (6) strand length.

produced by the alignment of proteins of known structure to all sequences considered homologous on the basis of a threshold curve, which is strongly dependent on the length of the alignment. HSSP sequence alignments (Sander and Schneider 1991) were incorporated into our method for some of the different strand attributes, as described in Materials and Methods. The inclusion of evolutionary information in this way improved the prediction of edge/central strands by 0.75%. This improvement is marginal, in comparison with the improvement in secondary structure prediction by the use of evolutionary information. The difference in the degree of conservation between secondary structure elements and loops accounts for a large proportion of the improvement in secondary structure prediction on the use of evolutionary information. Our results indicate that any difference in conservation between central and edge strands adds little to the prediction accuracy of our methods.

Interestingly, the rules generated automatically by the decision tree method, shown in Figure 3, reveal intuitive patterns that fit well with expected edge and central strand characteristics. The decision tree in Figure 3 represents a typical example from a single cross-validation experiment. Although all the trees generated contained similar patterns, they differed in minor details. For example, the tree shown in Figure 3 does not involve the hydrophobic moment, but this attribute appears in trees from different cross-validation data sets. The strand length attribute was common in all the decision trees, indicating that strand length is an important contributing factor to the final prediction. Figure 3 indicates that short strands (fewer than three residues) are representative of edge strands, supporting the findings of Sternberg and Thornton (1977). Strands of four residues in length are predicted to be central strands only if they are very hydrophobic (≤35.16% hydrophilic residues). For longer strands, the prediction involves more complex rules. Predominantly hydrophobic strands (≤43.75 hydrophilic residues) are still predicted central. However, even hydrophilic strands with a moderately large charge score (>0.0155) are predicted central if greater than five residues in length. If such charged hydrophilic strands are equal to five residues in length, then they are predicted to be edge if there is little periodicity of polarity (≤0.5); otherwise, they are predicted to be central

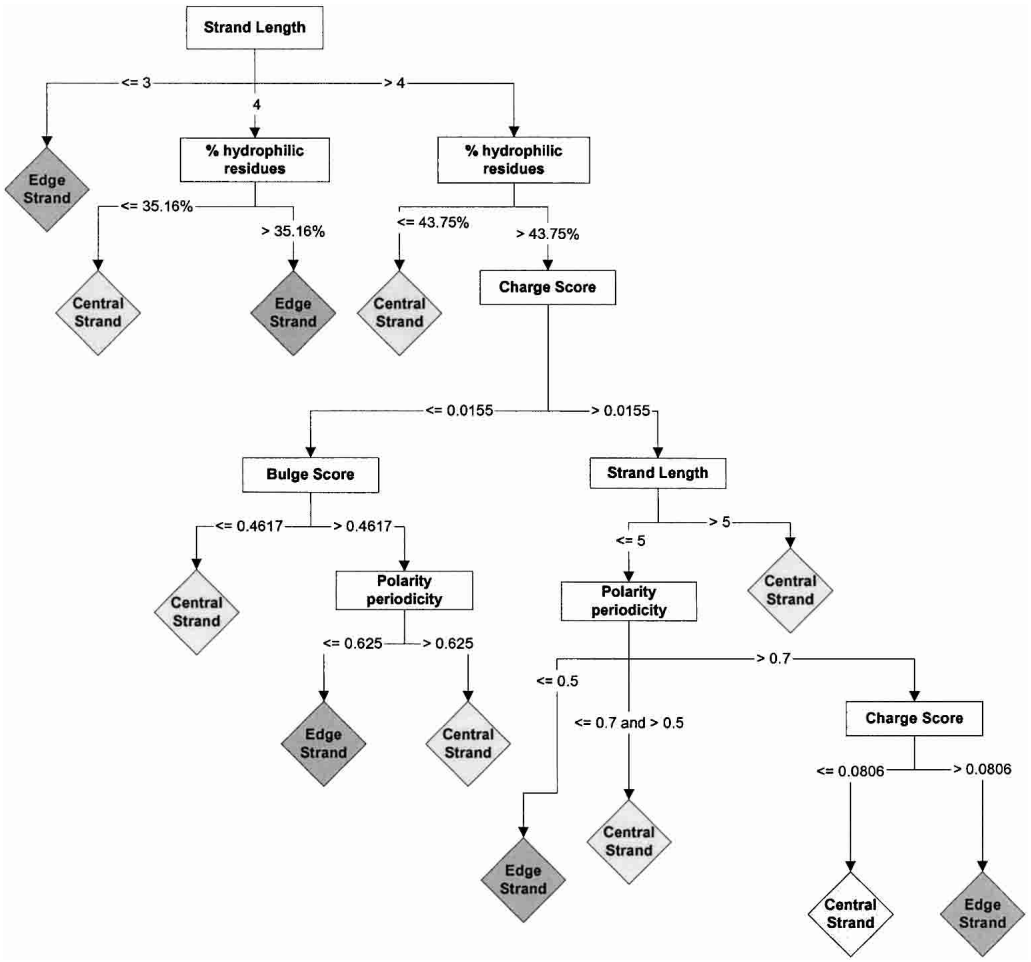

Figure 3. Decision tree. A typical decision tree, after pruning, from the C4.5 method.

unless they have very high charge scores (>0.0806). On the other hand, long hydrophilic strands with a low charge score (≤0.0155) are predicted to be edge if they have a high β bulge propensity (>0.4167) and low periodicity of polarity (≤0.625).

Investigation of a selection of mispredicted strands revealed some interesting limitations of the prediction method. RASMOL (Sayle and Milner-White 1995) and the Protein Quaternary Structure (PQS) file server (Henrick and Thornton 1998) were used to examine the location of a random selection of 100 mispredicted edge and central strands. The results are shown in Table 2. The prediction methods described here are based on the detection of edge strand features that protect against further interaction and aggregation. Prediction based on primary sequence means that only features intrinsic to the strand in question can be used. We found that ~50% of edge strands, mispredicted to be central strands, appear to be protected from further interactions and aggregation by extrinsic structural mechanisms, not solely involving the amino acid residues of the edge strand. For instance, five of the mispredicted edge strands were observed to lie at the core of β-propeller structures, which have four to eight radial blades of up-and-down β-sheets (Richardson and Richardson 2002). An example of this is the edge strand (residues 158–165) of the six-bladed propeller from the thermostable phytase protein. As shown in Figure 4A, this edge strand is completely protected at the propeller center, and consequently, no β edge protection strategy is present in the amino acid sequence of this strand.

The β edge protection strategy in 29 of edge strands mispredicted as central strands (Table 2) involved extrinsic protection from structures such as loops and helices of the same monomer, as well as the formation of higher oligomers. An edge strand (residues 48–51, chain A) of the RANTES (regulated upon activation, normal T-cell expressed and secreted) protein (Fig. 4B) is mispredicted as a central strand by the SVM method. As shown in Figure 4B, a large loop (residues 12–16) protects the edge strand from hydrogen bonding with any other β-strands. This loop coverage, which allows the β-sheet structure to remain regular and still be protected by a separate part of the polypeptide chain, has been described as a protective mechanism in other protein structures (Richardson and Richardson 2002). Similar effects are seen in the formation of higher oligomers, for which the PQS server (Henrick and Thornton 1998) predicts the protein to form part of a higher oligomeric structure than the representation of the protein structure by the coordinates deposited in the Protein Data Bank (PDB). For example, the mispredicted edge strand (residues 60–65) of the ACMNPV telokin-like protein appears to be unprotected from aggregation (Fig. 4C), indicating it should represent a typical β edge strand in our method. The PDB entry confers only a monomeric structure; however, the same protein is predicted to form part of a trimer by the PQS server (Henrick and Thornton 1998). As shown in Figure 4C, the unprotected mispredicted edge strand in the monomeric structure of the protein is protected by its position at the center of the trimer, even though formation of the trimer does not extend the β-sheet by hydrogen bonding from the mispredicted strand. Edge strands found in β-sheet multimers such as this are described by Richardson and Richardson (2002) to have less than one intrinsic protective feature per strand (compared with 2.5 features in comparable monomers), which is likely to result in a β edge strand with amino acid features more typical of a central strand.

As we commented above, strand length seems to be one of the most important attributes in our prediction method. Nevertheless, the investigations of mispredictions reported in Table 2 reveal that seven edge strands mispredicted as central strands consisted of more than five residues, which is a strand length more typical of a central strand, and that 19 central strands mispredicted as edge strands were equal to or less than three residues in length, with eight of those containing only one or two residues, a length more typical of an edge strand. Clearly strand length is an important indicator in many cases, but the derived rules do break down in some.

Table 2. *Analysis of a random selection of 100 incorrect SVM edge and central strand predictions of structurally assigned strands*

Possible explanation for misprediction	Strand assignment by DSSP[a]	Strand prediction by SVM method	No. of strands	Example
In the center of a propeller structure	Edge	Central	5	Fig. 4A
Helix/loop cover from the same monomer	Edge	Central	15	Fig. 4B
Helix/loop cover from other monomer (predicted by PQS[b])	Edge	Central	14	Fig. 4C
Long strand (more than five residues)	Edge	Central	7	—
Irregularity in the strand	Edge	Central	14	Fig. 4D
Unknown mechanism	Edge	Central	15	—
Short strand (three or less residues)	Central	Edge	19	—
Fewer than 50% of residues in the strand make hydrogen bonds to the adjacent strand	Central	Edge	7	Fig. 4E
Unknown mechanism	Central	Edge	4	—

[a] Kabsch and Sander (1983).
[b] Henrick and Thornton (1998).

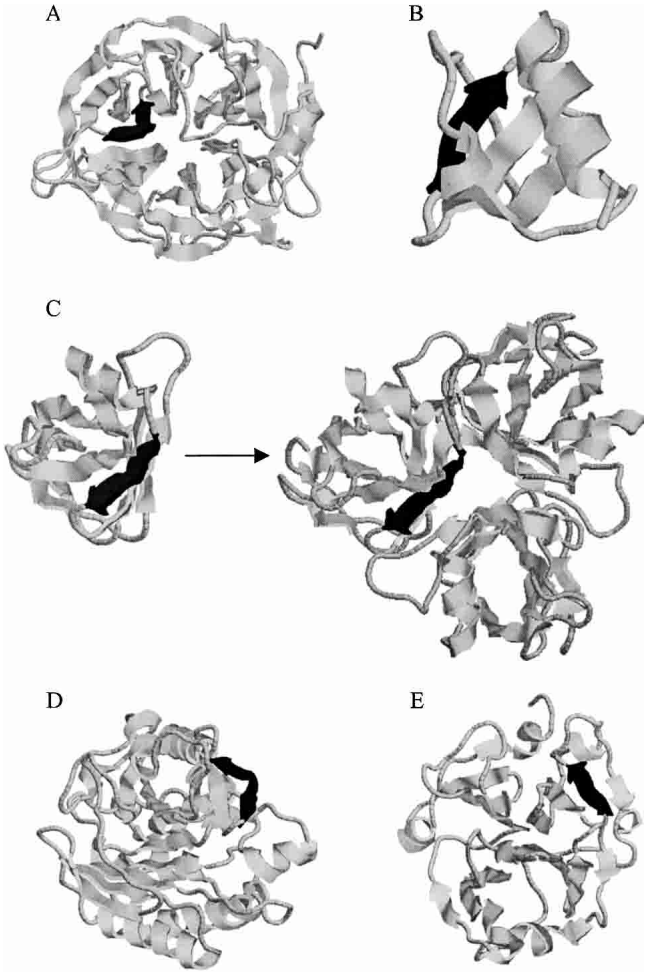

Figure 4. Examples of incorrect predictions made by the SVM method. All figures were produced by using RASMOL (Sayle and Milner-White 1995). (*A*) The six-bladed β-propeller of the thermostable phytase protein (1POO). The mispredicted edge strand (black) is located in the core of the propeller. (*B*) The monomeric RANTES protein (1B3A), with the mispredicted edge strand (black) protected by a covering loop. (*C*) The ACMNPV telokin-like protein (1TUL), with a β-clip fold, is monomeric in the PDB/RASMOL (Sayle and Milner-White 1995) structure (*left*). In PQS (Henrick and Thornton 1998), the same protein is predicted to form a trimer (*right*), where the mispredicted edge strand (black) is found in the center of the trimeric structure. (*D*) Phosphatidylinositol-specific phospholipase C (2PTD) with a TIM β/α-barrel fold. The mispredicted edge strand (black) contains a twist. (*E*) The four-bladed β-propeller structure of hemopexin protein (1HXN). The mispredicted central strand (black) has a short adjacent strand, resulting in more than half of the residues in the strand being exposed.

Structural irregularities feature in 14 examples of edge strands mispredicted as central strands. For example, a mispredicted edge strand of the phosphatidylinositol-specific phospholipase C protein, as shown in Figure 4D, contains a twist that is likely to disrupt or even prevent further hydrogen bonding to other β-strands. This is almost certainly a protective mechanism and has been observed in the edge strands of several β-sandwich proteins (Nesloney and Kelly 1996). But, in contrast to the prediction of β bulge irregularities, which is included in our method, the presence of the twist is difficult to predict from sequence and, therefore, difficult to incorporate.

Investigation of central strands mispredicted as edge strands also gave some interesting results. As shown in Table 2, seven central strands mispredicted as edge strands had partial edge strand character, owing to incomplete satisfaction of hydrogen bonds by short or twisted adjacent strands. For example, the central strand (residues 253–256) mispredicted as an edge strand from the four-bladed β-propeller structure of hemopexin protein (Fig. 4E) has a short adjacent edge strand. The adjacent strand contains only half the number of residues as the mispredicted strand.

Predictions based on predicted secondary structure

The results above indicate that given perfect (i.e., DSSP) knowledge of the positions of β-strands in a protein sequence, edge strands can be distinguished with reasonable accuracy from central strands. Here we consider the effect of using predicted secondary structure.

The profile network prediction Heidelberg (PHD) secondary structure prediction method (Rost and Sander 1993) was applied to the data set of proteins of known structure used above. By using the DSSP information, the PHD strand predictions can be further classified as correctly predicted central or edge strands, or as mispredicted strands (parts of the sequence incorrectly predicted to be strands). In this context, we define a strand as correctly predicted if a PHD predicted strand overlaps the DSSP strand by at least one residue. This generous criterion means that correctly predicted strands may differ significantly in their precise boundaries (first and last residues) from the DSSP assignments. The secondary structure predictions for this data set contain 1404 edge strands, 2685 central strands, and 971 mispredicted strands. The Q_{strand} scores (see Materials and Methods) for PHD (Rost and Sander 1993) predicted edge and central strands shown in Figure 5, clearly demonstrating that the prediction of edge strands is significantly worse than central strands (the difference in accuracy is significant at the 95% level in a Wilcoxon test), with <50% (1404 of 3359) of edge strands predicted correctly compared with ~90% (2685 of 2995) of central strands. Comparable results were obtained with the alternative PSIPRED (Jones 1999) method (data not shown).

The prediction method above, trained on DSSP-defined strands, was applied to the PHD strand predictions (note that in this case repeat cross-validations were not performed, so standard error was not available). A number of issues must be considered in evaluating performance in this case. First, PHD only detects ~50% of edge strands, effectively limiting the combined (PHD plus edge/central prediction) method to a significantly smaller number of strands in this category. Second, the PHD mispredicted strands (above) cannot be

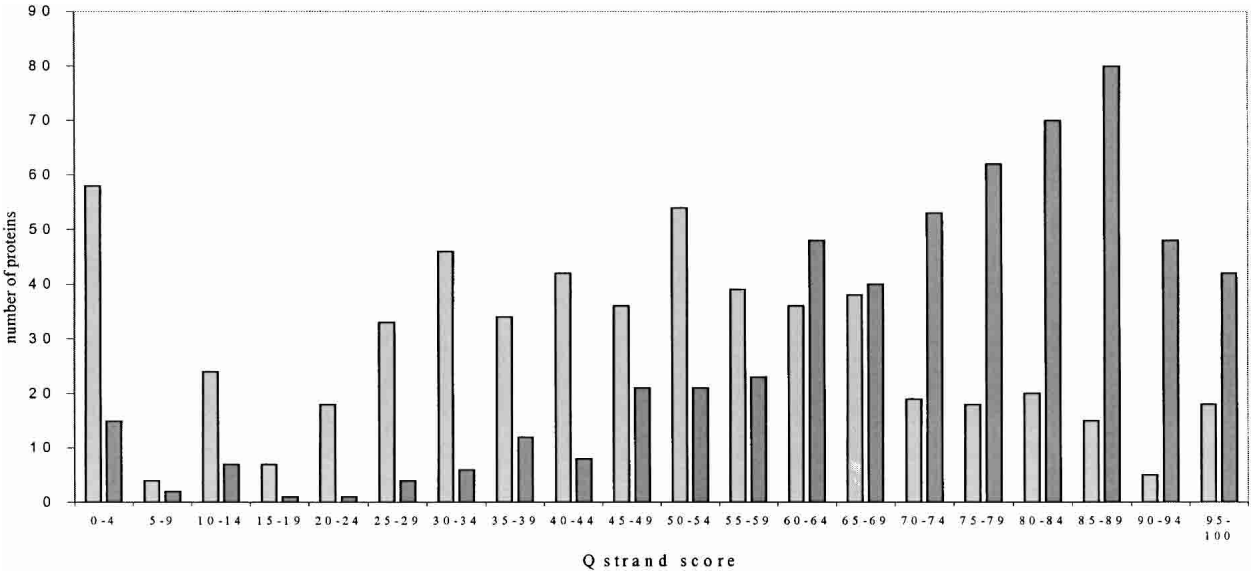

Figure 5. Edge and central strand prediction accuracy. Q_{strand} scores (correctly predicted strands) of PHD (Rost and Sander 1993) predicted edge and central strands. Bars shown in light gray represent the edge strands; bars in darker gray, the central strands.

classed as either edge or central, and so, any prediction our edge/central classifier might make for these strands should probably be considered an error. To provide a realistic measure of the expected performance of the combined method when applied to proteins of unknown structure, we adopt this latter approach and consider all PHD mispredicted strands as errors. However, in trying to understand the effect of predicted secondary structure on our method, it is also useful to look at performance indicators at which the PHD mispredicted strands are omitted. This limits the test data to strand predictions that are shared with the DSSP strand assignments, allowing a direct comparison of the prediction performance between actual and predicted secondary structure.

The results for predictions based on PHD strand assignments are shown in Table 3. First, considering the overall performance of the edge/central prediction method, it is clear that most performance measures are significantly lower when based on predicted secondary structure. The overall percentage of correct assignments is reduced from 78% to 55% for both learning methods, and in particular, the sensitivity and specificity of edge strand prediction are lowered to values in the range of 30% to 40%. Correspondingly, the Matthews coefficients are also much lower (~0.0). Nevertheless, this performance is still significantly better than either of the naive classifiers discussed previously (note that the approach of considering all PHD mispredicted strands as errors in edge/central prediction reduces the expected accuracy of the naive random classifier to 40%). However, despite this statistical significance, the poor performance of the combined method in predicting edge strands unfortunately means that it is probably of rather limited use in a practical context.

When the PHD mispredicted strands are omitted from the calculations (numbers in parentheses in Table 3) the apparent performance of the method obviously improves; however, most indicators are still lower than their values for predictions based on actual secondary structure by significant amounts (e.g., the overall percentage of correct assignments decreases from 78% to 69%). This reduction in performance can be attributed to two different sources: (1) the effect of imperfect predictions of the strand boundaries (adding or omitting residues on the ends of strands), and (2) a bias in the type of edge strands correctly predicted by PHD. The second effect is the most significant. The 1404 edge strands correctly predicted by PHD are split in the ratio 51:49 between strands that were correctly classified as edge based on DSSP strand assignments and strands that were wrongly classified. If the PHD predictions were randomly distributed between these two classes, the expected split would be closer to 80:20 (because the sensitivity [Table 1] of the method for edge strand prediction from DSSP data is 86% [SVM] or 74% [decision tree]). The correctly predicted edge strands from PHD are a set biased toward those that are difficult for our edge/central classifier, and this reduces the performance of our method when based on PHD predictions. This result is perhaps not unexpected: Edge strands are often different from central strands, but PHD predicts them better when they most resemble central strands and when they are most difficult to distinguish from central strands by our methods.

The bias above undoubtedly accounts for the largest part of the reduced performance of our method when based on predicted secondary structure. However, the effect of poor prediction of strand boundaries probably does affect accu-

Table 3. *The prediction accuracy of PHD-predicted (Rost and Sander 1993) edge and central strands*

Predicted strand class	Actual strand class: Central strands	Actual strand class: Edge strands	PHD mispredicted strands	PHD mispredicted strands included[a] (not included[b])
SVM method				
Central strands	2264	867	308	66% (72%)
Edge strands	421	538	662	33% (56%)
	84%	38%		55% (69%)
	Sensitivity			Specificity
Decision tree method				
Central strand	2220	856	298	66% (72%)
Edge strands	465	549	672	33% (54%)
	83%	39%		55% (68%)
	Sensitivity			

[a] Strands mispredicted by PHD are included in the test data and any prediction for them is assumed to be incorrect.
[b] Strands mispredicted by PHD are omitted from the test data.

racy to some extent. The attribute that would be affected most significantly is almost certainly the strand length, because all the other attributes are calculated as average values over the entire length of the strand and will, on average, be less sensitive to end effects. A full strand-by-strand analysis of the reasons for reduced performance on prediction data is impractical, but our concern over the accuracy of strand length predictions led us to investigate further. The results in Figure 2 indicate we can predict edge/central strands with reasonable accuracy without using strand length as an attribute. Using such a method and applying it to predicted secondary structure (data not shown) resulted in a similar reduction in performance as was found when strand length was included. Even on predicted structure, the method using strand length was still the more accurate. Strand length predictions might not be perfect, but they are good enough to have some predictive value in this context.

In addition to the methods described above, an alternative approach was taken to try and overcome the limitations imposed by the quality of PHD strand predictions. The method was trained on PHD-defined strands, but this performed worse both in overall accuracy and the sensitivity and specificity of edge strand prediction than did the same method trained on DSSP-defined strands (data not shown).

Discussion

Our results show that mechanisms used by naturally occurring proteins to protect against aggregation of β-sheets (Richardson and Richardson 2002) can be exploited to distinguish the edge strands from central strands on the basis of primary sequence. Both the machine learning methods used achieved accuracies ~78% of correct predictions. The SVM learning method would normally be considered to be more sophisticated than the decision tree and was expected to provide better generalization performance. However, in this case with our relatively small and carefully chosen set of predictive attributes, the performances of the methods were very similar. The decision tree method has the advantage of automatically generating interpretable rules, and in this case, those rules were in close accord with our knowledge of protein structure. For instance, shorter strands or longer more hydrophilic strands with reduced amphiphilic nature are predicted to be edge strands. Analysis of mispredictions revealed that edge strands mispredicted as central strands often lack protection intrinsic in their own sequence and structure, but rather are protected by extrinsic mechanisms involving other parts of the tertiary or quaternary structure. Central strands mispredicted as edge strands often had partial edge character with some exposed hydrogen bonding potential. It is interesting that our results show only minor improvement on the inclusion of evolutionary information from multiple alignments, indicating that the sequence of evolutionary related β-strands provides little additional information to distinguish edge and central strands.

The effect of using predicted secondary structure (PHD) rather than actual secondary structure (DSSP) on the performance of our methods was disappointing. Although the methods still performed significantly better overall than did naive prediction methods, the relatively poor performance in edge strand predictions means that the application of our method, in combination with predicted secondary structure, is probably currently of limited use in practical contexts, such as identifying edge strand residues likely to be involved in aggregation of proteins of unknown structure. Nevertheless, our efforts to understand the reasons for this disappointing performance revealed some effects that are of general interest in secondary structure prediction. Secondary structure prediction is very significantly less accurate for edge strands then for central strands. Further, edge strands that are missed by secondary structure prediction are predominantly those that have aggregation protection mechanisms intrinsic to their own sequences and structure, whereas those that are detected tend to share more properties with central strands and to be protected by extrinsic structural mechanisms. The edge strands detected by secondary structure prediction are those most difficult for our methods to distinguish from central strands.

Finally, despite the problems of using predicted secondary structure, the methods we present here are potentially useful in other contexts. For instance, our decision trees

provide a set of rules that can be used in the design of edge strands. These may be useful in preventing aggregation of de novo designed proteins, but equally, they could make a valuable contribution to the understanding and design of edge strand mutations to promote or inhibit the aggregation of naturally occurring proteins.

Materials and methods

The data set

Representative protein domains were extracted from the Structural Classification of Proteins (SCOP) database (Hubbard et al. 1997), release 1.55, by using the ASTRAL compendium (Brenner et al. 2000). For rigorous evaluation of machine learning methods, it is important to use a nonredundant data set. To ensure this, the representative protein domain from ASTRAL was taken for each superfamily in SCOP containing substantial β-structure (only protein domains from the β, α+β, α/β, and multidomain protein classes were used). Using superfamily representatives ensures that, according to SCOP, the data set does not contain any homologous pairs of proteins. In a small number of cases, it will contain protein pairs of the same fold for which the expert-derived SCOP classification cannot find strong evidence of homology. Overall, there were 564 proteins in the data set.

Assignment of edge and central strands

The DSSP algorithm (Kabsch and Sander 1983) was applied to all the representative domains in the database. Edge strands were defined as strands in which all residues have only a single bridge partner (see the DSSP article for the definition of a bridge partner), leaving one side of the strand available for hydrogen bonding; central strands were defined as strands in which at least one residue has bridge partners on both sides. Overall, there were 3359 edge strands and 2995 central strands in the data set.

Secondary structure prediction

The secondary structure prediction method, PHD (Rost and Sander 1993), was applied to the representative domains. Similar results were also obtained by using the PSIPRED method (Jones 1999; data not shown).

The Q_{strand} scores for both the edge and central strand predictions were calculated as the percentage of correctly predicted strands, at which if at least one residue from a particular strand was correctly predicted, the entire strand was classed as correctly predicted.

Machine learning

Prediction of edge strands from central strands based on sequence was achieved by using machine learning methods. Two main approaches were used: SVMs and decision trees.

SVM implementation

The application of SVMs in molecular biology is starting to gain significant interest, particularly as it frequently demonstrates high prediction accuracy. SVMs are becoming increasingly popular, mainly because of their ability to generalize well (avoid overfitting) but also because they can handle large feature spaces and condense the information given by the training data set (by use of support vectors; Hua and Sun, 2001). Molecular biology applications include protein solvent accessibility prediction (Yuan et al. 2002), protein secondary structure prediction (Hua and Sun 2001), gene expression classification (Brown et al. 2000), protein classification (Zavaljevski et al. 2002), and protein fold recognition (Ding 2001), whereas SVMs are comparable to or outperform other methods such as neural networks.

SVMs are a family of learning algorithms that find the optimal separating hyperplane from a set of binary-labeled data. The optimization algorithm, based on statistical learning theory (Vapnik 1998), maximizes the separating margin between the two classes of the training data and is defined by a small number of training data points called the support vectors (Cristianini and Shawe-Taylor 2000). For many problems in which the data cannot be separated in the input space by a linear function, kernel functions may be used to map the input space into a higher-dimensional feature space, enabling linear decision boundaries in the feature space to represent nonlinear decision boundaries in the input space (Cristianini and Shawe-Taylor 2000). SVMs were implemented by using SVMTorch (Collobert and Bengio 2001), which is freely downloadable from http://old-www.idiap.ch/learning/SVMTorch.html.

Decision trees

C4.5 (version 8; Quinlan 1993) was used to derive decision trees. C4.5 divides large sets of cases of both nominal and numerical properties belonging to known classes by identifying patterns within the cases (Quinlan 1993). These patterns are then expressed as models, in the form of decision trees, that can be used to classify new cases. A decision tree has either a leaf, indicating a class or a decision node that specifies some test to be carried out on a single attribute value. To avoid the construction of a complicated tree (overfitting on the data), the tree is pruned. Pruning usually discards one or more subtrees and replaces them with leaves at which the class is chosen by examining the training cases and choosing the most frequent class. From a decision tree, sets of if/then rules can be generated by generalization of each leaf of the tree (Quinlan 1993). C4.5 is freely downloadable from http://www.cse.unsw.edu.au/~quinlan.

Feature attribute extraction

Features, extracted from the amino acid sequence of each β-strand, were used as input to each machine learning method. For each strand in the data set, attributes were extracted from the primary sequence based on six characteristics thought to be dissimilar in the two strand types, based on published data discussed previously.

β bulges have been described as an important protective β edge strategy in β-propellers, β-sandwiches, and single β-sheets. β bulges are an irregularity found in β-strands in which two residues of the edge strand are opposite a single residue on the neighboring strand. This bulge in the strand makes it very difficult, if not impossible, to continue hydrogen bonding on the convex side of the strand; consequently, β bulges are unfavored in the interior of a β-sheet but desirable in the edges of a β-sheet. Chan et al. (1993) investigated the amino acid preferences for the five types of β bulge. These preferences were used to assign a bulge score (b_i) for each amino acid based on the number of occurrences of each

residue in all the different types of β bulges. For each β-strand, the total bulge score (B) was defined as

$$B = \frac{1}{L}\sum_{i=1}^{L} b_i$$

where L is the length of the strand, and b_i is the bulge score for residue i.

Strands within β-sheets show some hydrophobic ordering, with the most hydrophobic strand buried in the sheet and the other strands occurring in order of decreasing hydrophobicity, with the edge strands generally more hydrophilic. The proportion of hydrophilic residues (H) in each strand was used to model this effect using

$$H = \frac{1}{L}\sum_{i=1}^{L} h_i$$

where each residue was assigned a numerical hydrophobicity (h_i) of zero for hydrophobic residues (Ala, Gly, Val, Leu, Met, Trp, Cys, Pro, Phe, and Ile) and one for hydrophilic residues. In addition to hydrophobic ordering, it is well documented that edge strands are generally shorter than more centrally located strands. The length of each strand was also used as input to the machine learning methods.

β-sheets often have an amphiphilic nature, with patterns of alternating hydrophobic and hydrophilic residues causing one side of the β-sheet to be hydrophobic and the other hydrophilic, which may be disrupted by the protective β edge features of the edge strands. Eisenberg et al. (1984) defined the hydrophobic moment for the detection of the strength of periodic components, and it was applied in our method to assign a numerical value to the periodicity of hydrophobic/hydrophilic residues in each of the β-strands. Each residue was assigned a numerical hydrophobicity of one for hydrophobic residues (Ala, Gly, Val, Leu, Met, Trp, Cys, Pro, Phe, and Ile) and −1 for hydrophilic residues. The hydrophobic moment (Hm) was defined as (based on the method of Eisenberg et al. 1984).

$$Hm = \frac{1}{L}\left|\sum_{i=1}^{L} h_i(-1)^{i-1}\right|$$

In addition, an alternative measure of periodicity in polarity was also used. This is defined by

$$P = \frac{1}{L}\sum_{i=1}^{L} A_i$$

where $A_i = 1$ if residues i and $i + 1$ are of opposite polarity and zero otherwise. This measure is a count of the number of polarity swaps along the length of each strand.

Finally, the positioning of charged residues in the edge strands is a common β edge protection strategy, particularly prevalent in β-propellers and single β-sheets. The charge score (C) was defined as

$$C = \frac{1}{L}\sum_{i=1}^{L} c_i$$

where the charge (c_i) was one for each charged amino acid (Asp, Glu, Lys, Arg, His) and zero for all other amino acid residues.

To investigate the effect of including evolutionary information in the prediction method, the above quantities were averaged over a set of HSSP (Sander and Schneider 1991) aligned strand sequences.

Parameter optimization

The SVM parameters were optimized by using the 18-fold cross-validation technique with the structurally assigned strand information. Some of the most successful results, with different kernels and parameters are shown in Table 4. SVMTorch (Collobert and Bengio 2001) has two other kernel functions, the sigmoidal and polynomial kernels, both of which performed considerably worse than did the Gaussian kernel on the data (data not shown). The Gaussian kernel with the parameter set to 6 SD outperformed the other parameters and kernel functions and was used in the remainder of the analysis. The default parameters for the decision tree method were found to be approximately optimal and were used throughout.

Measuring performance

To increase the readability of the article, we describe our cross-validation techniques and performance measures at the beginning of the Results section. Performance measures include the overall percentage of correct predictions (compared with naive prediction methods by using a one-sample t test), sensitivity, specificity, and MCC (Matthews 1975). The latter is defined, for each class i (edge/central), as

$$C_i = \frac{p_i n_i - u_i o_i}{\sqrt{(p_i + o_i)(p_i + u_i)(n_i + o_i)(n_i + u_i)}}$$

where p_i is the number of correctly predicted strands of type i (edge, central, or mispredicted strands), n_i is the number of strands correctly not assigned to type i, u_i is the number of underestimated

Table 4. *Example of parameters used during optimization of the kernel and parameters for the SVM method*

Kernel	Parameter	Edge strands predicted correctly	Central strands predicted correctly	Total edge/central strands predicted correctly
Linear[a]	—	75%	77%	76%
Gaussian[b]	Std. 2	79%	74%	77%
	Std. 6	72%	86%	79%
	Std. 9	78%	76%	77%
	Std. 11	73%	74%	74%

[a] $(a,b) \mapsto (a*b)$

[b] $(a,b) \mapsto \exp\left(\frac{-\|a-b\|^2}{std^2}\right)$

strand types, o_i the number of overestimated strand types, and C_i is MCC for a particular class, for two class problems $C_1 = C_2$. MCC is a single measure of performance, accounting for both under- and overpredictions and has been used to assess the prediction accuracy of the location and type of β-turns in protein sequences (Shepherd et al. 1999). Finally, for each prediction, the sensitivity and specificity of the prediction of edge and central strands were calculated. The sensitivity is the proportion of the central/edge strands predicted to be central/edge strands respectively. Specificity is the proportion of strands predicted to be central/edge strands that actually are central/edge strands.

Acknowledgments

We would like to thank the Leeds Bioinformatics group for useful discussions and the MRC for sponsorship.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

References

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

Brenner, S.E., Koehl, P., and Levitt, M. 2000. The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res.* **28:** 254–256.

Brown, M., Grundy, W., Lin, D., Cristianini, N., Sugnet, C., Ares, M., and Haussler, D. 2000. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci.* **97:** 262–267.

Bystroff, C. and Baker, D. 1998. Prediction of local structure in proteins using a library of sequence–structure motifs. *J. Mol. Biol.* **281:** 565–577.

Chan, A.W., Hutchinson, E.G., Harris, D., and Thornton, J.M. 1993. Identification, classification, and analysis of β-bulges in proteins. *Protein Sci.* **2:** 1574–1590.

Collobert, R. and Bengio, S. 2001. SVMTorch: Support vector machines for large-scale regression problems. *J. Machine Learning Res.* **1:** 143–160.

Cristianini, N. and Shawe-Taylor, J. 2000. *Support vector machines and other kernel-based learning methods*. Cambridge University Press, Cambridge, MA.

Ding, C.H. and Dubchak, I. 2001. Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics* **17:** 349–358.

Eisenberg, D., Weiss, R.M., and Terwilliger, T.C. 1984. The hydrophobic moment detects periodicity in protein hydrophobicity. *Proc. Natl. Acad. Sci.* **81:** 140–144.

Henrick, K. and Thornton, J.M. 1998. PQS: A protein quaternary structure file server. *Trends Biochem. Sci.* **23:** 358–361.

Hill, C.P., Yee, J., Selsted, M.E., and Eisenberg, D. 1991. Crystal structure of Defensin HNP-3, an amphiphilic dimer: Mechanisms of membrane permeabilization. *Science* **251:** 1481–1485.

Hoshino, M., Katou, H., Hagihara, Y., Hasegawa, K., Naiki, H., and Goto, Y. 2002. Mapping the core of the β2-microglobulin amyloid fibril by H/D exchange. *Nat. Struct. Biol.* **9:** 332–336.

Hua, S. and Sun, Z. 2001. A novel method of protein secondary structure prediction with high segment overlap measure: Support vector machine approach. *J. Mol. Biol.* **308:** 397–407.

Hubbard, T.J., Murzin, A.G., Brenner, S.E., and Chothia, C. 1997. SCOP: A structural classification of proteins database. *Nucleic Acids Res.* **25:** 236–239.

Hutchinson, E.G., Sessions, R.B., Thornton, J.M., and Woolfson, D.N. 1998. Determinants of strand register in antiparallel β-sheets of proteins. *Protein Sci.* **7:** 2287–2300.

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

Kabsch, W. and Sander, C. 1983. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22:** 2577–2637.

Kelley, L.A., MacCallum, R.M., and Sternberg, M.J. 2000. Enhanced genome annotation using structural profiles in the program 3DPSSM. *J. Mol. Biol.* **299:** 501–522.

King, R.D., Clark, D.A., Shirazi, J., and Sternberg, M.J. 1994. On the use of machine learning to identify topological rules in the packing of β-strands. *Protein Eng.* **7:** 1295–1303.

Lopez de la Paz, M., Goldie, K., Zurdo, J., Lacroix, E., Dobson, C.M., Hoenger, A., and Serrano, L. 2002. De novo designed peptide-based amyloid fibrils. *Proc. Natl. Acad. Sci.* **99:** 16052–16057.

Matthews, B.W. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta.* **405:** 442–451.

Minor, D. and Kim, P. 1994. Context is a major determinant of β-sheet propensity. *Nature* **371:** 264–267.

Monti, M., Principe, S., Giorgetti, S., Mangione, P., Merlini, G., Clark, A., Bellotti, V., Amoresano, A., and Pucci, P. 2002. Topological investigation of amyloid fibrils obtained from β2-microglobulin. *Protein Sci.* **11:** 2362–2369.

Nesloney, C.L. and Kelly, J.W. 1996. Progress towards understanding β-sheet structure. *Bioorg. Med. Chem.* **4:** 739–766.

Pollastri, G., Przybylski, D., Rost, B., and Baldi, P. 2002. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins* **47:** 228–235.

Przybylski, D. and Rost, B. 2002. Alignments grow, secondary structure prediction improves. *Proteins* **46:** 197–205.

Quinlan, J.R. 1993. *C4.5: Programs for machine learning*. Morgan Kaufmann Publishers, San Mateo, CA.

Richardson, J.S. and Richardson, D.C. 2002. Natural β-sheet proteins use negative design to avoid edge-to-edge aggregation. *Proc. Natl. Acad. Sci.* **99:** 2754–2759.

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

Sander, C. and Schneider, R. 1991. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins* **9:** 56–68.

Sayle, R. and Milner-White, J. 1995. RasMol: Biomolecular graphics for all. *Trends Biochem. Sci.* **20:** 374.

Serag, A.A., Altenbach, C., Gingery, M., Hubbell, W.L., and Yeates, T.O. 2001. Identification of a subunit interface in transthyretin amyloid fibrils: Evidence for self-assembly from oligomeric building blocks. *Biochemistry* **40:** 9089–9096.

———. 2002. Arrangement of subunits and ordering of β strands in an amyloid sheet. *Nat. Struct. Biol.* **9:** 734–739.

Serpell, L.C., Goldsteins, G., Dacklin, I., Lundgren, E., and Balke, C. 1996. The "edge strand" hypothesis: Prediction and test of a mutational hotspot on the transthyretin molecule associated with FAP amyloidosis. *Amyloid* **3:** 75–85.

Shepherd, A.J., Gorse, D., and Thornton, J.M. 1999. Prediction of the location and type of β-turns in proteins using neural networks. *Protein Sci.* **8:** 1045–1055.

Simons, K.T., Kooperberg, C., Huang, E., and Baker, D. 1997. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *J. Mol. Biol.* **268:** 209–225.

Sternberg, M.J. and Thornton, J.M. 1977. On the conformation of proteins: Hydrophobic ordering of strands in β-pleated sheets. *J. Mol. Biol.* **115:** 1–17.

Steward, R.E. and Thornton, J.M. 2002. Prediction of strand pairing in antiparallel and parallel β sheets using information theory. *Proteins* **48:** 178–191.

Trinh, C.H., Smith, D.P., Kalverda, A.P., Philips, S.E., and Radford, S.E. 2002. Crystal structure of monomeric human β2-microglobulin reveals clues to its amyloidogenic properties. *Proc. Natl. Acad. Sci.* **99:** 9771–9776.

Vapnik, V. 1998. *Statistical learning theory.* Wiley, New York.

Wang, W. and Hecht, M.H. 2002. Rationally designed mutations convert de novo amyloid-like fibrils into monomeric β-sheet proteins. *Proc. Natl. Acad. Sci.* **99:** 2760–2765.

Yuan, Z., Burrage, K., and Mattick, J.S. 2002. Prediction of protein solvent accessibility using support vector machines. *Proteins* **48:** 566–570.

Zaremba, S.M. and Gregoret, L.M. 1999. Context-dependence of amino acid residue pairing in antiparallel β-sheets. *J. Mol. Biol.* **291:** 463–479.

Zavaljevski, N., Stevens, F.J., and Reifman, J. 2002. Support vector machines with selective kernel scaling for protein classification and identification of key amino acid positions. *Bioinformatics* **18:** 689–696.

Zvelebil, M.J., Barton, G.J., Taylor, W.R., and Sternberg, M.J. 1987. Prediction of protein secondary structure and active sites using the alignment of homologous sequences. *J. Mol. Biol.* **195:** 957–961.