

Protein Topology Classification Using Two-Stage Support Vector Machines

Jayavardhana Gubbi¹

jrg1@ee.unimelb.edu.au

Alistair Shilton¹

apsh@ee.unimelb.edu.au

Michael Parker²

mparker@svi.edu.au

Marimuthu Palaniswami¹

swami@ee.unimelb.edu.au

¹ Department of Electrical and Electronics Engineering, The University of Melbourne, Parkville, Victoria 3010, Australia

² St. Vincent's Institute of Medical Research, 9 Princes Street, Fitzroy, Victoria 3065, Australia

Abstract

The determination of the first 3-D model of a protein from its sequence alone is a non-trivial problem. The first 3-D model is the key to the molecular replacement method of solving phase problem in x-ray crystallography. If the sequence identity is more than 30%, homology modelling can be used to determine the correct topology (as defined by CATH) or fold (as defined by SCOP). If the sequence identity is less than 25%, however, the task is very challenging. In this paper we address the topology classification of proteins with sequence identity of less than 25%. The input information to the system is amino acid sequence, the predicted secondary structure and the predicted real value relative solvent accessibility. A two stage support vector machine (SVM) approach is proposed for classifying the sequences to three different structural classes (α , β , $\alpha + \beta$) in the first stage and 39 topologies in the second stage. The method is evaluated using a newly curated dataset from CATH with maximum pairwise sequence identity less than 25%. An impressive overall accuracy of 87.44% and 83.15% is reported for class and topology prediction, respectively. In the class prediction stage, a sensitivity of 0.77 and a specificity of 0.91 is obtained. Data file, SVM implementation (SVMHEAVY) and result files can be downloaded from <http://www.ee.unimelb.edu.au/ISSNIP/downloads/>.

Keywords: protein topology classification, kernel methods, support vector machines

1 Introduction

X-ray crystallography is one of the most popular experimental methods for determining the three dimensional structure of proteins. Indeed, more than 80% of the structures deposited into the protein data bank (PDB) [2] have been obtained using this method. One of the challenges when using this method is solving the phase problem [9]. A popular approach to this problem is the Molecular Replacement (MR) method [33], which is motivated by the observation that if the structures of two proteins are identical, then the diffraction patterns produced by them should be identical [9]. Based on this observation, the MR searches the PDB to find the protein therein which has the maximum sequence identity with the unknown protein. This known protein is called the first model. The phase of the first model is then combined with the intensity information of the unknown protein, and together these form a starting point for the calculation of the actual phase (and hence the three-dimensional structure) of the unknown protein. This latter process usually involves the determination of the positions of the model within the crystal cell using repeated rotation and translation functions [9, 33]. It is known [38] that if the protein sequence identity of the two proteins is greater than 35% then the proteins have similar structure. This knowledge is used in homology modeling to determine the

suitability of the first model. However this becomes challenging when the sequence identity is less than 35% (and even more so if the sequence identity is less than 25%) [38].

Three unique methods are available to define protein structure for experimentally determined models: FSSP (Fold classification based on Structure-Structure alignment of Proteins) [15], SCOP (Structural Classification of Proteins) [32] and CATH (Class, Architecture, Topology and Homology) [36]. FSSP uses Dali [16] for structure-structure alignment of proteins. SCOP divides the proteins into four hierarchical classes - Family, Superfamily, Fold and Class. FSSP and SCOP both use evolutionary relationships for classification. CATH is another type of hierarchical classification which uses the SSAP (Sequential Structure Alignment Program) [36] algorithm and is based on structural comparison. We choose to use CATH definitions and database for all the work reported in this paper as we found it to be more suitable for our future work.

Fold recognition as defined by SCOP is nearest to topology as defined by CATH. Hence we review the literature for fold recognition followed by topology prediction methods. FORESST [7] uses secondary structure information and hidden Markov models (HMM) for fold recognition. Ding and Dubchak [8, 10] combined support vector machines (SVMs) and neural networks for fold recognition. They tested their method on 27 SCOP folds with less than 25% sequence identity. Tan *et al.* [43] used an ensemble machine learning approach for the same problem, and showed that their scheme was effective in scenarios where there are multi-class unbalanced datasets. Rangwala and Karypis [37] used SVMs with two novel kernels for fold recognition and remote homology detection. They used several profile based features in their work. Lund *et al.* [28] proposed a method to construct the sequence profile for a sequence for which a ready template is not available. Fold and Function Assignment System (FFAS03) [18] tries to match profiles obtained from PSI-BLAST for fold recognition. 3DSHOTGUN [12, 13] is a meta-predictor which uses results from other major fold recognition schemes. It was rated in the top three servers in CAFASP3 experiments. SAM-T02 [24] uses HMM for protein structure prediction. This server has also proved to be very effective and is rated among the top servers. GenThreader [29] is another popular technique which uses feed forward neural network and predicted secondary structure for structure prediction.

Relative to fold recognition, not much work has been done in the field of topology prediction. This is due to the fact that the classification method used by SCOP is based on evolutionary information, which is said to be more reliable than CATH. However, due to the convenience CATH provides for our future work, we have decided to use topology classification based on CATH. Early work by Di Francesco *et al.* [6] used HMM for topology prediction from secondary structure, but can only be used on alpha class proteins. By contrast, in this paper we propose a method which first identifies the class (α , β , $\alpha + \beta$) and then classifies between 39 topologies. A lot of work has been done toward predicting only the structural class (a nice review of which can be found in [5]). The most recent paper in structural class prediction was by Cao *et al.* [4] and made use of the dataset by Zhou *et al.* [47]. The maximum pairwise sequence identity of the Zhou datasets (taken from SCOP-1997) is not clearly indicated and hence it is difficult to compare this with our data. Our data is from a more recent version of CATH (2005) and is a hard dataset with maximum pairwise sequence identity less than 22% for 99.8% of the data. Furthermore, our work involves the more complex problem of topology prediction and hence it is different from class prediction methods. More recently Lo *et al.* [27] used an SVM classifier for transmembrane helix and topology prediction. However the topology they define is entirely different to the CATH definitions that we use.

In this paper, we propose a new method of predicting class and topology as defined by CATH [36]. We achieve this by using a two stage SVM. We develop a structure alignment kernel based on dynamic programming, predicted secondary structure and Chou-Fasman conformational parameters in the first stage. In the second stage, we use predicted solvent accessibility and evolutionary information in the form of position specific scoring matrix (PSSM) obtained from PSI-BLAST in an intuitive way and show that these simple features can produce some excellent results.

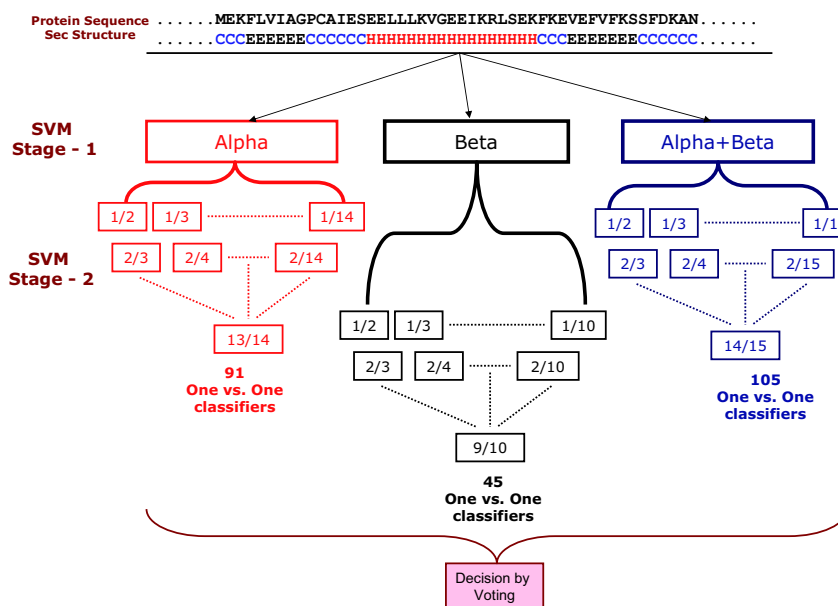


Figure 1: Proposed classification scheme.

2 Material and Methods

We have constructed our dataset (which we call it GSP742) using CATH version 2.6.0 (released in April 2005). This dataset can be downloaded from <http://www.ee.unimelb.edu.au/ISSNIP/downloads/>. To eliminate identical sequences, we have applied UniqueProt [31] with an HSSP-value of 5 to our dataset to eliminate identical sequences. We have retained only proteins with sequence length greater than 60 and resolution of at least 2 Å in our dataset. After doing this, we are left with 946 proteins out of 10,000+ which have pairwise identity of less than 30%. Finally, we have subjected all sequences to a pairwise global alignment algorithm to eliminate sequences with more than 25% identity. After this procedure, we are left with 742 sequences with a maximum pairwise sequence identity of less than 25%, of which less than 10 (0.01%) fall into the range 22%-25%. There are 164 proteins belonging to class α , 223 proteins belonging to class β and 355 proteins belonging to class $\alpha + \beta$. Of the total 39 topologies, there are 14 topologies in class α , 10 topologies in class β and 15 topologies in class $\alpha + \beta$. Our aim is to classify these proteins into three classes in first stage and 39 topologies in second stage.

Our method makes use of two layers of support vector machines, as shown in Figure 1. The system input is the protein sequence and its predicted secondary structure. At the first level, the protein sequences are classified into three classes as defined by CATH, namely α , β and $\alpha + \beta$ (which we will sometimes denote as γ). We do not make use of the fourth class, *few secondary structures*, as there are insufficient examples available for topologies in this class (with sequence identity less than 20%) to design a classifier. In the second stage, several binary classifiers are employed to differentiate 39 topologies (refer Table 2) and the best match is selected using voting. In the absence of first stage of classification, to classify T topologies, $\frac{1}{2}T(T-1)$ binary classifiers are required. In our case, 741 classifiers would be required, which is infeasible. By the introduction of multi level classification, we are able to reduce the number of classifiers to $3 + 91 + 45 + 105 = 244$ (3 classes in first stage followed by 14 α , 10 β and 15 γ topologies). We make use of predicted secondary structures [19] and predicted real value solvent accessibility [20] developed by our group. The prediction accuracy of the secondary structure method we use is in the range of 72% for unseen data. The cross-validation accuracy is 77%.

To predict solvent accessibility, we have employed adaptive support vector regression [20]. The Mean Absolute Error (MAE) of solvent accessibility prediction using cross validation is approximately 0.12. The 8 to 3 state reduction method of secondary structure [23] used was H, G and I to H; E and B to E; and all others to C (where H stands for Alpha Helix, E for Beta Strand and C stands for Coil).

Feature Extraction

We extract Chou-Fasman conformational parameters separately for α class (class 1), β class (class 2) and $\alpha + \beta$ class (class 3). The Chou-Fasman parameter for a helix (H) in class 1 is given by $P1_{Hi} = f1_{Hi}/\langle f1_H \rangle$ where $\langle f1_H \rangle$ is the number of residues in the helix divided by the total number of residues; and the index i selects from the set of 20 amino acids residues. Similar conformational parameters for strand $P1_{Ei}$ and coil $P1_{Ci}$ were calculated for class 1. Hence $P1$ is a 3×20 matrix with rows representing secondary structure states and columns representing amino acid residues. This notation is used in our new kernel development. The procedure is repeated for class 2 (P2) and class 3 (P3). This will be used for features for first stage of classification using the structure matching kernel described later.

The folding free energy can be expressed as the summation of free energies due to intra molecular interaction and the interaction with the surrounding solvent molecules [35]. Most solvation models assume that the solvation energy of the solute is the sum of individual solvation energies of the residues. This would also give an indication about the position of the residue with respect to the core of protein which will enable the calculation of accessible surface area of a residue [17, 35]. These values (relative solvent accessibility - RSA) reflect the contribution of each side chain to the thermodynamic parameters of hydration which give an indication of hydrophobicity and hydrophilicity. We use predicted real values of RSA [20] as one of our features.

We also extracted evolutionary information in the form of a position specific scoring matrix (PSSM) generated by PSI-BLAST [1] using the non-redundant (NR) database. The low complexity regions, coiled-coil regions and transmembrane helices were filtered with *pfilt* [22]. We chose an E-value of 0.0001 and 10 iterations for PSI-BLAST. The BLOSUM62 matrix was used for multiple sequence alignment.

The RSA and evolutionary information extracted are used in the second level of classification. To ensure that the length of all feature vectors is constant irrespective of sequence length, we represent it in a unique way. The hydrophobicity scale so extracted is converted to a feature vector by taking the sum of the relative solvent accessibility (RSA) values for each amino acid. This contributes 20 dimensions to each feature vector. Similarly for evolutionary information, we calculate the sum of PSSM values for each amino acid for which the values are greater than 0, which contributes 20 more dimensions, giving a total of 40 dimensions for each feature vector. Intuitively, these give some indication of the contribution of each amino acid to the hydrophobicity scale and the evolutionary scale.

3 Support Vector Machines

Support Vector Machines introduced by Vapnik [45] has become one of the most popular tools in bioinformatics for supervised classification. These are binary classification algorithms based on structural risk minimization [3, 42, 46]. SVMs do this by implicitly mapping the training data into a (usually higher-dimensional) feature space. A hyperplane (decision surface) is then constructed in this feature space that bisects the two categories and maximises the margin of separation between itself and those points lying nearest to it (the support vectors). This decision surface can then be used as a basis for classifying vectors of unknown classification. The final decision function is given by [39, 45]:

$$g(\mathbf{y}) = \sum_{(\mathbf{x}_i, d_i) \in \Theta} \alpha_i d_i K(\mathbf{x}_i, \mathbf{y}) + b.$$

The function $K(\mathbf{x}_i, \mathbf{x}_j) = \varphi(\mathbf{x}_i)^T \varphi(\mathbf{x}_j)$ is called the kernel function, and plays a vital role in the SVM classifier. This is because the kernel function K completely hides the feature map $\varphi : \mathfrak{R}^{d_L} \rightarrow \mathfrak{R}^{d_H}$. Moreover, it is well known that for any function $K : \mathfrak{R}^{d_L} \times \mathfrak{R}^{d_L} \rightarrow \mathfrak{R}$ satisfying Mercer's condition [14, 30, 39] there exists an associated feature map $\varphi : \mathfrak{R}^{d_L} \rightarrow \mathfrak{R}^{d_H}$ (although calculating this map may not be a nontrivial exercise). Hence starting with any such kernel function we may, with no knowledge of the feature map φ at all, optimize and use an SV-classifier based on that kernel function (and with no knowledge of the feature map concealed therein). This is referred to as the *Kernel trick* [39]. Mercer's condition states that $K : \mathfrak{R}^{d_L} \times \mathfrak{R}^{d_L} \rightarrow \mathfrak{R}$ must be a continuous, non-negative definite, symmetric kernel function.

4 Structure Alignment Kernel

In the first stage of our proposed classifier, we make use of a dynamic programming (DP) kernel [41] for aligning the structure. In [41], a dynamic time warping kernel was developed for speech recognition. We have used the same basic procedure for sequence alignment, but with a significantly different process for constructing the DP matrix (D) and the final decision function. A few other sequence alignment kernels developed can be found in [11, 25, 26, 40]. We have tried two methods for building the DP matrix:

Using Chou-Fasman conformational parameters: For the α model, we first extract Chou-Fasman parameters as described in section 2. Let the target sequence (Tar) be of length M and the template sequence (Tem) be of length N . We define the predicted secondary structure states H, E, C of the target and template sequence as $TarSS$ and $TemSS$, respectively. Our input matrix (In) and DP matrix (D) are of size $M \times N$. The Input Matrix In is constructed using $P1$ via the following function:

$$In(i, j) = |P1(TarSS_i, Tar_i) - P1(TemSS_j, Tem_j)|, \quad (1)$$

where $TarSS_i$ and $TemSS_j$ are target and template secondary structure states of residues i and j ; and Tar_i and Tem_i the target and template amino acids. From the input matrix, we calculate the DP matrix D using:

$$D(i, j) = \min \left\{ \begin{array}{l} D(i-1, j) + In(i, j) \\ D(i-1, j-1) + 2In(i, j) \\ D(i, j-1) + In(i, j) \end{array} \right\}. \quad (2)$$

As can be seen from the above equation, $D(M, N)$ is the least cumulative cost between the template and target sequences. If the same sequence is given as template and target, the final cost will be zero. Thus $D(M, N)$ represents the difference, or *distance*, between the vectorial representations of two sequences. A Mercer kernel K , on the other hand, represents an *inner product* between the vectorial representation of two sequences. To construct an inner product from a distance measure, we use the same trick used in the RBF kernel, namely:

$$K(Tar, Tem) = 100 \exp \left(\frac{-D(M, N)}{\delta(M + N)} \right), \quad (3)$$

where δ was chosen (experimentally) to be 0.01 for all our experiments. Essentially the same procedure is repeated for β model and γ models.

Using an empirically chosen substitution matrix: Three substitution matrices (S) were empirically chosen for modelling the three classes thusly:

$$S_\alpha = \begin{bmatrix} 4 & -2 & -2 \\ -2 & 1 & -2 \\ -2 & -2 & 2 \end{bmatrix}, \quad S_\beta = \begin{bmatrix} 1 & -2 & -2 \\ -2 & 4 & -2 \\ -2 & -2 & 2 \end{bmatrix}, \quad S_\gamma = \begin{bmatrix} 4 & -2 & -2 \\ -2 & 4 & -2 \\ -2 & -2 & 2 \end{bmatrix}. \quad (4)$$

From the above matrices, it is clear that any mismatch in secondary structure will get a score of -2 . Considering only S_H , substituting H by H (correct match) will get high score of 4 as the model under consideration is alpha. Similarly substituting E by E will get 1 and *coil* with *coil* will get 2. Again, we develop the method for α model and extend it to other models. We use the same notations as described earlier. The input Matrix In is constructed using S_H via the following function:

$$In(i, j) = S_\alpha(TarSS_i, TemSS_j), \quad (5)$$

where $TarSS_i$ and $TemSS_j$ are the target and template secondary structure states of residues i and j . From the input matrix, we calculate the DP matrix D using:

$$D(i, j) = \max \left\{ \begin{array}{l} D(i-1, j) + Gap \\ D(i-1, j-1) + 2In(i, j) \\ D(i, j-1) + Gap \end{array} \right\}, \quad (6)$$

where Gap in the above equation is gap penalty. In all of our experiments we have used $Gap = -10$. The kernel function for this method is defined as follows:

$$K(Tar, Tem) = \frac{D(M, N)}{(M + N)}. \quad (7)$$

The kernel thus designed may not be positive definite. We use empirical kernel map to convert this to a valid kernel [34, 44]. Both the methods gave good results, with neither having a clear advantage over the other. However, as the latter method proved to be less computationally expensive, we have chosen to use it for the remainder of our paper. For the three classifiers in the first stage, the constant C which controls the trade-off between the dual objectives of maximising the margin of separation and minimising the misclassification error was chosen to be 1, 1 and 0.1, respectively (values were chosen experimentally).

For all the classifiers in the second stage of classification, we used radial basis function (RBF) kernel (eq. 8) with $\gamma = 0.01$; and C were empirically chosen to be 10. The RBF kernel is as follows:

$$K(x, y) = \exp \left(\frac{-\|x - y\|^2}{\gamma} \right). \quad (8)$$

5 Results and Discussion

We use standard accuracy (Q), sensitivity and specificity to evaluate our result in the first stage. These are defined to be:

$$Q = \frac{TP + TN}{TP + TN + FP + FN}, \quad Sensitivity = \frac{TP}{(TP + FN)}, \quad Specificity = \frac{TN}{(FP + TN)}, \quad (9)$$

where TP = True Positive, TN = True Negative, FP = False Positive, and FN = False Negative. 10 fold cross validation was used in the first stage and jackknife testing was used in the second stage to evaluate our method. The results are summarized in Table 1. As it can be seen from this table, the results are impressive with high sensitivity and specificity. The overall accuracy of class prediction was about 87.44%, and topology prediction was 83.15%. The result for $\alpha + \beta$ appears surprising, with topology accuracy greater than class accuracy. The reason for this is that, in case where no class is a winner in the first stage, we subject the test protein to all one vs. one classifiers in the topology stage. As the result of high accuracy in topology state, there is an increase the final result. Table 2 summarises the accuracies obtained for each topology. The combined accuracies in Table 2 represent the output of the entire system; and the topology accuracies corresponds to the accuracy of the stand alone topology prediction (assuming 100% recognition by the class predictor). Out of 39 topologies

Table 1: Result of jackknife test on GSP742 dataset.

Model	Sensitivity	Specificity	Class Accuracy	Topology Accuracy
α	0.75	0.97	92.45%	74.99%
β	0.76	0.92	87.20%	79.37%
$\alpha + \beta$	0.79	0.88	83.29%	89.29%
Overall	0.77	0.91	87.44	83.15%

6 (15%) scored less than 50% in the jackknife test. This was mainly due to the error in recognizing the class in the first stage. The stand alone accuracy for the same topologies is 100%. In particular, Glycosyltransferase (1.50.10) was always getting recognized as belonging to $\alpha + \beta$ instead of class α which resulted in an extremely low value.

6 Conclusion

In this paper we have developed a classifier based on support vector machines for the classification of Class and Topology of proteins as defined by CATH. We have developed a new structure alignment kernel which has been shown to work impressively even with a secondary structure prediction rate in early seventies. Using a better secondary structure prediction method such as PSIPRED [21] may produce even better results. The solvent accessibility information and evolutionary information has been combined for topology prediction and the stand alone accuracy of topology classification for most topologies shown to be nearly 100%. Overall accuracy of 87.44% and 83.15% has been obtained for class and topology prediction, respectively. In the class prediction stage, we have reported an overall sensitivity of 0.77 and an overall specificity of 0.91.

Table 2: Topologies used in classification with class, CAT code, number of protein chains in each topology, combined accuracies and topology accuracies.

Class	CAT Code	Topology Name	Number of Protein Chains		Combined Accuracies (%)		Topology Accuracies (%)	
			Protein Chains	Accuracies (%)	Protein Chains	Accuracies (%)		
α	1.10.10	Arc Repressor Mutant, subunit A	25	52.00	100.00	100.00	100.00	
α	1.10.150	DNA polymerase; domain 1	8	75.00	100.00	100.00	100.00	
α	1.10.20	Histone, subunit A	10	100.00	100.00	100.00	100.00	
α	1.10.238	Recoverin; domain 1	14	85.71	100.00	100.00	100.00	
α	1.10.287	Helix Hairpins	12	83.33	100.00	100.00	100.00	
α	1.10.490	Globin-like	15	100.00	100.00	100.00	100.00	
α	1.10.760	Cytochrome Bcl Complex; Chain D, domain 2	10	60.00	100.00	100.00	100.00	
α	1.10.8	Helicase, Ruva Protein; domain 3	8	25.00	100.00	100.00	100.00	
α	1.20.120	Four Helix Bundle (Hemerythrin (Met), subunit A)	17	94.12	100.00	100.00	100.00	
α	1.20.1250	Growth Hormone; Chain A;	9	88.89	100.00	100.00	100.00	
α	1.20.5	Single alpha-helices involved in coiled-coils or other helix-helix interfaces	10	80.00	100.00	100.00	100.00	
α	1.20.58	Methane Monooxygenase Hydroxylase; Chain G, domain 1	7	100.00	100.00	100.00	100.00	
α	1.25.40	Serine Threonine Protein Phosphatase 5, Tetratricopeptide repeat	13	76.92	100.00	100.00	100.00	
α	1.50.10	Glycosyltransferase	6	0.00	100.00	100.00	100.00	
β	2.10.60	CD59	9	33.33	100.00	100.00	100.00	
β	2.30.29	PH-domain like	7	42.86	100.00	100.00	100.00	
β	2.30.30	SH3 type barrels.	19	63.16	100.00	100.00	100.00	
β	2.30.42	Pdz3 Domain	6	50.00	100.00	100.00	100.00	
β	2.40.10	Thrombin, subunit H	11	90.91	100.00	100.00	100.00	
β	2.40.128	Lipocalin	13	100.00	100.00	100.00	100.00	
β	2.40.50	OB fold (Dihydroipoamide Acetyltransferase, E2P)	30	40.00	100.00	100.00	100.00	
β	2.60.120	Jelly Rolls	42	90.48	100.00	100.00	100.00	
β	2.60.40	Immunoglobulin-like	80	96.25	98.75	100.00	100.00	
β	2.80.10	Trefoil (Acidic Fibroblast Growth Factor, subunit A)	6	100.00	100.00	100.00	100.00	
$\alpha + \beta$	3.10.129	Thiol Ester Dehydrase; Chain A	8	37.50	100.00	100.00	100.00	
$\alpha + \beta$	3.10.180	2,3-Dihydroxybiphenyl 1,2-Dioxygenase; domain 1	8	87.50	100.00	100.00	100.00	
$\alpha + \beta$	3.10.20	Ubiquitin-like (UB roll)	15	100.00	100.00	100.00	100.00	
$\alpha + \beta$	3.10.450	Nuclear Transport Factor 2; Chain A,	14	42.86	100.00	100.00	100.00	
$\alpha + \beta$	3.20.20	Bactericidal permeability-increasing protein; domain 2	49	95.92	95.92	100.00	95.92	
$\alpha + \beta$	3.30.360	Dihydrodipicolinate Reductase; domain 2	6	100.00	100.00	100.00	100.00	
$\alpha + \beta$	3.30.420	Nucleotidyltransferase; domain 5	6	83.33	100.00	100.00	100.00	
$\alpha + \beta$	3.30.70	Alpha-Beta Plaits	40	95.00	100.00	100.00	100.00	
$\alpha + \beta$	3.40.190	D-Maltodextrin-Binding Protein; domain 2	11	100.00	100.00	100.00	100.00	
$\alpha + \beta$	3.40.30	Glutaredoxin	15	100.00	100.00	100.00	100.00	
$\alpha + \beta$	3.40.50	Rossmann fold	151	88.74	90.73	100.00	90.73	
$\alpha + \beta$	3.40.630	Aminopeptidase	11	100.00	100.00	100.00	100.00	
$\alpha + \beta$	3.40.640	Aspartate Aminotransferase; domain 2	5	100.00	100.00	100.00	100.00	
$\alpha + \beta$	3.60.20	Glutamine Phosphoribosylpyrophosphate, subunit 1, domain 1	7	85.71	100.00	100.00	100.00	
$\alpha + \beta$	3.90.550	Spore Coat Polysaccharide Biosynthesis Protein SpsA; Chain A	9	88.89	100.00	100.00	100.00	

References

- [1] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J., Gapped BLAST and PSI-BLAST: A new generation of protein database search programs, *Nucleic Acids Res.*, 27(17):3389–3402, 1997.
- [2] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E., The protein data bank, *Nucleic Acids Res.*, 28(1):235–242, 2000.
- [3] Burges, C.J.C., A tutorial on support vector machines for pattern recognition, *Knowledge Discovery and Data Mining*, 2(2):121–167, 1998.
- [4] Cao, Y., Liu, S., Zhang, L., Qin, J., and Wang, J., Prediction of protein structural class with rough sets, *BMC Bioinformatics*, 7(20), 2006.
- [5] Chou, K.C., Progress in protein structural class prediction and its impact to bioinformatics and proteomics, *Current Protein and Peptide Science*, 6(5):423–436, 2005.
- [6] Di Francesco, V., Garnier, J., and Munson, P.J., Protein topology recognition from secondary structure sequences: Application of the hidden Markov models to alpha class proteins, *J. Mol. Biol.*, 267:446–463, 1997.
- [7] Di Francesco, V., Munson, P.J., and Garnier J., FORESST: Fold recognition from secondary structure predictions of proteins, *Bioinformatics*, 15:131–140, 1999.
- [8] Ding, C.H.Q. and Dubchak, I., Multi-class protein fold recognition using support vector machines and neural networks, *Bioinformatics*, 17:349–358, 2001.
- [9] Drenth., J., *Principles of Protein X-Ray Crystallography - 2nd Edition*, Springer Verlag, 1999.
- [10] Dubchak, I., Muchnik, I., Mayor, C., Dralyuk, I., and Kim, S.H., Recognition of a protein fold in the context of the structural classification of proteins (SCOP) classification. *Proteins*, 35:401–407, 2005.
- [11] Eskin, E. and Snir, S., The homology kernel: A biologically motivated sequence embedding into euclidean space, *Proceedings of the 2005 IEEE Symposium on CIBCB'05*, 1–8, 2005.
- [12] Fischer, D., 3D-SHOTGUN: A novel, cooperative, fold-recognition meta-predictor, *Proteins*, 51:434–441, 2003.
- [13] Fischer, D., 3DS3 and 3DS5 3D-SHOTGUN meta-predictors in CAFASP3, *Proteins*, 53:517–523, 2003.
- [14] Haussler, D., Convolution kernels on discrete structures, *Technical Report*, UCS-CRL-99(10), 1999.
- [15] Holm, L. and Sander, C., The FSSP database of structurally aligned protein fold families, *Nucleic Acids Res.*, 22(17):3600–3609, 1994.
- [16] Holm, L. and Sander, C., Touring protein fold space with Dali/FSSP, *Nucleic Acids Res.*, 26:316–319, 1998.
- [17] Jaramillo, A. and Wodak, S.J., Computational protein design is a challenge for implicit solvation model, *Biophysical Journal*, 88:156–171, 2005.
- [18] Jaroszewski, L., Rychlewski, L., Li, Z., Li, W., and Godzik, A., FFAS03: A server for profile-profile sequence alignments, *Nucleic Acids Res.*, 33:W284–W288, 2005.

- [19] Jayavardhana Rama, G.L., Palaniswami, M., Lai, D., and Parker, M.W., A study on the effect of physico-chemical properties in protein secondary structure prediction, *Applied Artificial Intelligence*, 609–616, World Scientific, 2006.
- [20] Jayavardhana Rama, G.L., Shilton, A., Palaniswami, M., and Parker, M., Real value solvent accessibility prediction using adaptive support vector regression, *Department of EEE, The University of Melbourne, Technical Report*, 2006.
- [21] Jones, D.T., Protein secondary structure prediction based on position-specific scoring matrices, *J. Mol. Biol.*, 292:195–202, 1999.
- [22] Jones, D.T., Taylor, W.R., and Thornton, J.M., A model recognition approach to the prediction of all-helical membrane protein structure and topology, *Biochemistry*, 33:3038–3049, 1994.
- [23] Kabsch, W. and Sander, C., Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features, *Biopolymers*, 22:2577–2637, 1983.
- [24] Karplus, K., Barrett, C., and Hughey, R., Hidden Markov models for detecting remote protein homologies, *Bioinformatics*, 14:846–856, 1998.
- [25] Leslie, C., Eskin, E., Cohen, A., Weston, J., and Noble, W.S., Mismatch string kernels for discriminative protein classification, *Bioinformatics*, 20(4):467–476, 2004.
- [26] Leslie, C., Eskin, E., and Noble, W.S., The spectrum kernel: A string kernel for SVM protein classification, *Pac. Symp. Biocomput.*, 7:564–575, 2002.
- [27] Lo, A., Chiu, H.-S., Sung, T.-Y., and Hsu, W.-L., Transmembrane helix and topology prediction using hierarchical SVM classifiers and an alternating geometric scoring function. *Proceedings of IEEE Computational Systems Bioinformatics Conference (CSB06)*, 31–42, 2006.
- [28] Lund, O., Nielsen, M., Lundegaard, C., and Worning, P., CPHmodels 2.0: X3M a computer program to extract 3D models, *Abstract at the CASP5 conference*, A102, 2002.
- [29] McGuffin, L.J., Bryson, K., and Jones, D.T., The PSIPRED protein structure prediction server, *Bioinformatics*, 16:404–405, 2000.
- [30] Mercer, J., Functions of positive and negative type, and their connection with the theory of integral equations, *Transactions of the Royal Society of London (A)*, 209:415–446, 1909.
- [31] Mika, S. and Rost, B., UniqueProt: Creating representative protein-sequence sets, *Nucleic Acids Res.*, 31(13):3789–3791, 2003.
- [32] Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C., SCOP: A structural classification of proteins database for the investigation of sequences and structures, *J. Mol. Biol.*, 247:536–540, 1995.
- [33] Navaza, J., Implementation of molecular replacement in AMoRe, *Acta Crystallogr. D Biol. Crystallogr.*, 57:1367–1372, 2001.
- [34] Nobel, W.S., *Support Vector Machine Applications in Computational Biology*, MIT press, 2004.
- [35] Ooi, T., Oobatake, M., Nemethy, G., and Scheraga, H.A., Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides, *Proc. Natl. Acad. Sci. USA*, 84:3086–3090, 1987.
- [36] Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., and Thornton, J.M., CATH – a hierarchic classification of protein domain structures, *Structure*, 5(8):1093–1108, 1997.

- [37] Rangwala, H. and Karypis, G., Profile-based direct kernels for remote homology detection and fold recognition, *Bioinformatics*, 21:4239–4247, 2005.
- [38] Rost, B., Twilight zone of protein sequence alignments, *Protein Engineering*, 12(2):85–94, 1999.
- [39] Schölkopf, B., and Smola, A., *Learning with Kernels: Support Vector Machines, Regularization, Optimization and beyond*, MIT press, 2002.
- [40] Shawe-Taylor, J. and Cristianini, N., *Kernel Methods for Pattern Analysis*, Cambridge University press, 2004.
- [41] Shimodaira, H., Noma, K., Nakai, M., and Sagayama, S., Dynamic time-alignment kernel in support vector machine, *Advances in Neural Information Processing Systems 14*, MIT press, 2002.
- [42] Smola, A., and Schölkopf, B., A tutorial on support vector regression, *NeuroCOLT2 Technical Report Series*, NC2-TR-1998-030, Royal Holloway College, University of London, UK, 1998.
- [43] Tan, A.C., Gilbert, D., and Deville, Y., Multi-class protein fold classification using a new ensemble machine learning approach, *Genome Inform.*, 14:206–217, 2003.
- [44] Tsuda, K., Support vector classification with asymmetric kernel function, *Proceedings of the Seventh European Symposium on Artificial Neural Network*, 183–188, 2004.
- [45] Vapnik, V., *Statistical Learning Theory*, Springer-Verlag, New York, 1995.
- [46] Vapnik, V., Golowich, S., and Smola, A., Support vector methods for function approximation, regression estimation, and signal processing, *Advances in Neural Information Processing Systems*, MIT press, 1997.
- [47] Zhou, G.P., An intriguing controversy over protein structural class prediction, *Journal of Protein Chemistry*, 17(8):729–738, 1998.