A Hierarchical Approach to Protein Fold Prediction

Tabrez Anwar Shamim Mohammad[1] and Hampapathalu Adimurthy Nagarajaram[2]

Laboratory of Computational Biology, CDFD, Bldg.7, Gruhakalpa, Nampally, Hyderabad 500 001 India, http://www.cdfd.org.in

Summary

Fold recognition, assigning novel proteins to known structures, forms an important component of the overall protein structure discovery process. The available methods for protein fold recognition are limited by the low fold-coverage and/or low prediction accuracies. We describe here a new Support Vector Machine (SVM)-based method for protein fold prediction with high prediction accuracy and high fold-coverage. The new method of fold prediction with high fold-coverage was developed by training and testing on a large number of folds in order to make the method suitable for large scale fold predictions. However, presence of large number of folds in the training set made the classification task difficult as a consequence of increased complexity involved in binary classifications of SVMs. In order to overcome this complexity we adopted a hierarchical approach where fold-prediction is made in two steps. At the first step structural class of the query is predicted and at the second step fold is predicted within the predicted structural class. This decreased the complexity of the classification problem and also improved the overall fold prediction accuracy. To the best of our knowledge this is the first taxonomic fold recognition method to cover over 700 protein-folds and gives prediction accuracy of around 70% on a benchmark dataset. Since the new method gives rise to state of the art prediction performance and hence can be very useful for structural characterization of proteins discovered in various genomes.

1 Introduction

In this post-genomics era, there is a huge gap between the number of proteins with only sequence information and the number of proteins with both sequence and experimentally determined structural information. Since protein's tertiary (3D) structure provides important insights about its function, it is very important to have the knowledge of 3D structural information of all the proteins for the systems-level understanding of life. As all the proteins are not amenable for experimental structural determination there is a greater need for accurate computational prediction of protein structures than ever before. Protein fold recognition forms the most important component of the overall protein structure prediction problem and it refers to a method of assigning the most compatible structural fold out of the known structural folds for a given protein sequence. Protein fold recognition methods offer powerful and accurate means to detect structural homologues that are otherwise difficult to detect by conventional sequence-based homology search methods. A number of methods have been developed and used for protein fold recognition and these can be broadly grouped in three categories: (a) Sequence-structure homology recognition methods, (b) Threading methods and (c) Taxonomic methods. Among these approaches, the taxonomic methods, such as TAXFOLD [1], ACCFold [2], SVM-Fold [3], Shamim *et al*'s method [4], PFRES [5], PFP-pred [6], achieve highest prediction accuracies but are limited by the low fold-coverage. On the other

[1] Present address: Greehey Children's Cancer Research Institute, UTHSC, San Antonio, Texas, USA 78229; E-mail: mohammadt@uthscsa.edu

[2] Corresponding author. E-mail: han@cdfd.org.in

doi:10.2390/biecoll-jib-2011-185

hand, first two kinds of fold recognition methods like 3DPSSM/Phyre [7], FUGUE [8] & THREADER [9] have high fold-coverage, but their prediction accuracies are very low in comparison to taxonomic-based fold recognition methods. Therefore, it is imperative to explore the ways to develop a new taxonomic-based method for protein fold recognition with high prediction accuracy as well as high fold-coverage.

Earlier, we had developed a new taxonomic-based protein fold recognition method with the highest prediction accuracy by investigating fold-discriminatory potential of a number of protein sequence- and structure-based features [4]. The new method gave state of the art prediction performance on benchmark dataset [4]. However, the method was trained and tested on a benchmark dataset comprising of only 27 known protein folds referred to as D-B dataset [10] which is far less than the total number known protein folds which presently stands at 1195 [11]. Therefore, the method as such is not ready for large scale protein fold predictions. It is worthwhile to mention here that the other existing taxonomic fold-recognition methods too have been trained on small sets of known protein folds (Table 1) [1-6, 10, 12, 13] and it is to be noted that the respective authors have not made attempts to increase the number of protein folds despite having a potential for high prediction accuracies. This is because training of these tools necessarily involves at least 7-25 protein structures per fold and many folds do not have these many protein structures, and moreover, the task of dealing with large number of folds in a classification based set up is computationally quite challenging. Our method uses predicted structural information and therefore it is not essential to have all proteins with experimentally determined structural information. In the present study, we discuss how to increase the fold-coverage of the method by including more folds to the fold library and the eventual development of a novel hierarchical protein fold prediction method with the best prediction accuracy and high protein fold-coverage.

2 Methods

2.1 Dataset

Dataset for the studies was derived from the ASTRAL SCOP 1.73 database [14]. The protein sequences were extracted for four different pair-wise sequence identity cut-off values viz. <40%, <50%, <70% and <90% of the datasets from ASTRAL SCOP site (http://astral.berkeley.edu/). In the <40% dataset no two sequences have 40% or more than 40% sequence identity to each other. Proteins with short amino acid sequences and those having breaks were not included into any of the datasets.

2.1.1 711F Dataset

This dataset contains 33745 protein sequences belonging to 711 different SCOP folds. This includes 215 folds having ≥10 protein structures; 420 folds having 2-9 structures and 76 orphan folds having lone representative structure in ASTRAL SCOP 1.73 database. In order to populate folds having <10 protein structures, protein sequences related to representative protein/s were retrieved from sequence databases like Pfam [15] and only those folds were included to the fold library for which enough hits were found to make the cut-off of at least 10 proteins per fold.

Table 1: Fold-coverage (Number of protein folds covered) of different taxonomic fold recognition methods

Reference	Method	Fold-coverage (Number of Folds)
Yang and Chen, 2011	TAXFOLD	194
Kavousi *et al*., 2011	Kavousi *et al's* method	27
Dong *et al.*, 2009	ACCFold	199
Deschavanne and Tuffery, 2009	Deschavanne & Tuffery's method	60
Shamim *et al.*, 2007	Shamim *et al's* method	27
Melvin *et al.*, 2007	SVM-Fold	26
Chen and Kurgan, 2007	PFRES	27
Shen and Chou, 2006	PFP-pred	27
Ding and Dubchak, 2001	Ding and Dubchak's method	27

2.2 Classification Algorithm

We have shown earlier that SVM works well for protein fold prediction [4] and hence SVM was chosen for this study too. Since SVM has been primarily designed for binary classification [16] whereas protein fold prediction is a typical multi-class classification task and hence a multi-class method viz, *one versus one* was used in order to extend SVM into a multi-class classification task. *One-versus-one* method was chosen because we have shown earlier that one-versus-one method is computationally quicker than the other multi-class methods viz, *one versus all* and *Crammer & Singer* and also gives comparable accuracy to that of the other two methods [4]. *One versus one* multi-class method converts a multi-class problem into a series of all possible binary pair problems and uses a voting scheme to assign the most probable class to the query protein. All SVM computations were done using RBF kernel of LIBSVM [17] with optimized values of the cost parameter C and the kernel parameter γ. The optimization of SVM models were carried out by searching the optimum values of the cost parameter C = [2^{11}, 2^{10}… 2^{-3}] and kernel parameter γ = [2^{-11}, 2^{-10} … 2^{3}]. The optimized parameters were retained for further SVM training.

2.3 Features

The best feature from our earlier benchmarking studies [4] on protein fold classification was selected and used for training and testing. This feature is a combination of secondary structural state and burial state information of amino acids and amino acid pairs. These frequencies were calculated as follows.

2.3.1 Secondary structural state frequencies of amino acids and amino acid pairs

These are calculated using the formulae:

$$f_i^s = \frac{N_i^s}{L}$$

$$f\left(D_s^{i,i+n}\right)_j = \frac{N\left(D_s^{i,i+n}\right)_j}{L-n}$$

where $s = (H, E, C)$; H is α-helix; E is β-strands; C is ciols; $f^s{}_i$ is the frequency of amino acid i occurring in the secondary structural state s; and $N^s{}_i$ is the number of amino acid i found in the secondary structural state s; $f(D^{i,i+n}{}_s)_j$ is the frequencies of an n^{th} order amino acid pair j in secondary structural state s; and $N(D^{i,i+n}{}_s)_j$ is the number of an n^{th} order amino acid pair j found in secondary structural state s. These frequencies were calculated using predicted secondary structural information from PSIPRED [18] and as reported earlier by us only those with confidence level ≥1 were considered for calculations [4]. An amino acid pair was considered to be in helix or strand if both the residues are in helix or strand, respectively; rest considered as in coil.

2.3.2 Solvent accessibility state frequencies of amino acids and amino acid pairs

These are calculated using the formulae:

$$f_i^s = \frac{N_i^s}{L}$$

where $s = (B, E)$; $f^s{}_i$ is the frequency of amino acid i in solvent accessibility state s; and $N^s{}_i$ is the number of amino acid i in solvent accessibility state s. We have used predicted solvent accessibility states information derived from ACCpro [19] as the basis for all feature calculations and as reported earlier the cut off value for relative solvent accessibilities were ≤ 10% and >10% for buried (B) and exposed (E) respectively [4].

$$f\left(D_s^{i,i+n}\right)_j = \frac{N\left(D_s^{i,i+n}\right)_j}{L-n}$$

where $s = (B, E, I)$; $f(D^{i,i+n}{}_s)_j$ and $N(D^{i,i+n}{}_s)_j$ are the frequency and number of the n^{th} order amino acid pair j found in solvent accessibility state s; and L-n is the total number of n^{th} order amino acid pairs. An amino acid pair was considered as buried (B) or exposed (E) if both the residues were found buried or exposed, respectively; rest all the pairs were considered as partially buried (I).

2.4 Performance Measure

The performance of SVM fold-classifier was evaluated by computing overall accuracy (Q). The overall accuracy was calculated using the formula:

$$Q = \frac{\sum_i z_i}{N} \times 100$$

where N is the total number of proteins (instances) in the test set, and z_i are the true positives. The sensitivity and specificity values were calculated as reported earlier [4].

3 Results and Discussion

A look on the ASTAL SCOP 1.73 database shows that in total 79 folds hold true to the criteria where each fold contains at least 25 protein domains and none of the proteins in a fold share ≥40% sequence identity to each other (Figure 1). It may be noted here that the extended D-B dataset, which was used earlier for benchmarking purposes, comprises of only those protein folds, which has at least 25 protein domains and none of the proteins in the fold have ≥40% sequence identity to each other [4]. Further, a detailed survey of the ASTAL SCOP 1.73 database also revealed that the number of folds in the training and testing sets can be increased by: (a) increasing the pair-wise sequence identity of the dataset, (b) decreasing the cut-off for the minimum number of protein domains per fold (Figure 1). Moreover, since our

method uses predicted values of structural information; it can be extended further to include orphan folds (the folds with only one known protein structure are referred to as "orphan folds") and folds with very few protein structures by populating these folds with homologous sequences extracted from Pfam [15]. Orphan fold prediction is not possible by any other available method. It may be noted that orphan folds constitute about 20% of the total number of folds.

In order to increase the number of folds one, therefore, has to decrease the number of protein domains per fold or/and also increase the pair-wise sequence identity cut-off value of the dataset (Figure 1). Relaxing the criteria of pair-wise sequence identity cut-off value of the dataset from <40% to <90% and minimum number of protein domains in each fold from 25 to only one led to the increase in fold-coverage to 711. We included all the orphan folds and also those folds with less than ten protein structures to the fold library for which we got at least in total ten protein sequences including Pfam seed hits. This dataset is referred to as '711F' and contains 33745 protein sequences belonging to 711 different folds encompassing all four major structural classes (all-α, all-β, α/β & α+β). Distribution of protein folds in each structural class is listed in Table 2. However, before carrying out SVM learning on such a large dataset we studied the effect of increase in number of folds on the accuracy of prediction.

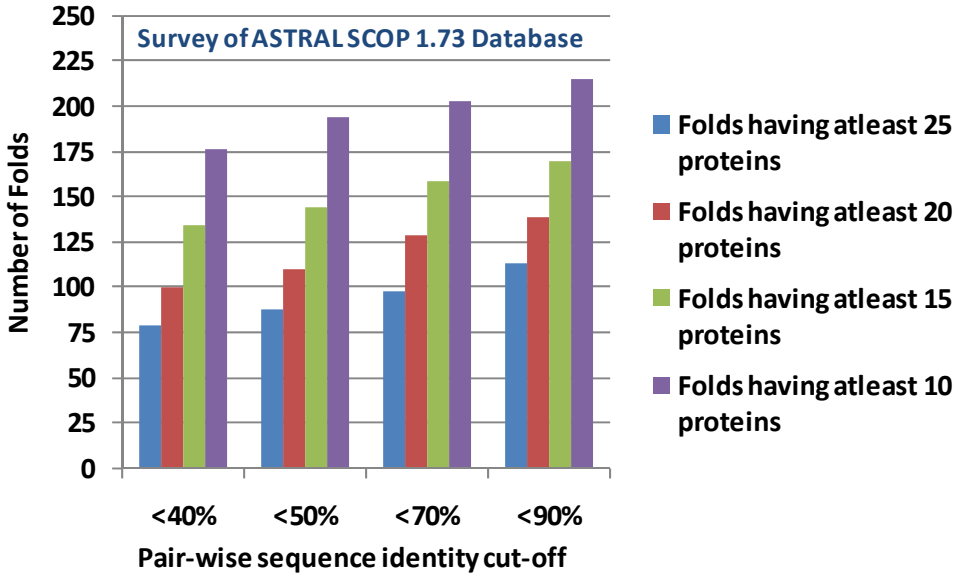

Figure 1: ***Survey of ASTRAL SCOP 1.73 database:*** **This figure depicts the total number of folds having ≥25 protein domains, ≥20 protein domains, ≥15 protein domains and ≥10 protein domains in ASTARL SCOP 1.73 database at 4 different sequence identity cut of values (<90%, <70%, <50% and <40%) of the dataset.**

Table 2: Distribution of protein folds in four major SCOP structural classes in three different datasets

	Number of Folds		
Structural Class	79-Folds dataset	176-Folds dataset	711F dataset
All-α Class	17	39	187
All-β Class	22	38	130
α/β Class	21	46	129
α+β Class	19	53	265

3.1 Effect of increased number of folds for training on protein fold prediction accuracy

The effect of increase in number of protein folds on prediction accuracies was studied using 79-Folds and 176-Folds dataset derived from the <40% ASTRAL SCOP 1.73 database. As evident from the Figure 1, <40% dataset has in total 79 folds with at least 25 proteins and 176 folds with at least 10 proteins, respectively. The protein sequences corresponding to these folds were extracted from ASTRAL SCOP 1.73 database and secondary structures and solvent accessibility states were predicted using PSIPRED [18] and ACCpro [19], respectively. The outputs of the PSIPRED and ACCpro were used for the calculation of discriminatory features as described earlier in Methods section. SVMs were trained using these fold-discriminatory features for both 79-Folds and 176-Folds dataset and training and testing of SVM models were studied by checking their fold-prediction accuracies. The overall five-fold cross-validation accuracies of 50% and 44% were obtained for 79 folds and 176 folds, respectively. The accuracy obtained for 79 and 176 folds classification is much less than 70% which we earlier obtained for 27 folds classification [4]. These results show that prediction accuracies decrease with the increase in the number of folds. This is probably because as the number of folds increases the number of pair-wise classifiers also increases exponentially and as a consequence the chance of identifying the correct fold becomes more ambiguous. The effect of large number of folds on prediction accuracy can be reduced to a certain extent if one can somehow split/convert a large fold classification task into a series of relatively smaller sets of fold classification tasks. The hierarchical nature of SCOP [20] and CATH [21] protein classification provides a way of reducing the effect of large classification problem. Both SCOP [20] and CATH [21] databases are hierarchical protein domain classification of protein structures. There are four major levels of hierarchy: Class, Architecture, Topology (fold family) and Homologous Superfamily in case of CATH [21] whereas SCOP comprises of: Class, Fold, Superfamily and Family [20]. Since we were interested in predicting SCOP folds; instead of directly predicting the fold of a protein, we adopted a hierarchical approach for protein fold recognition where first structural class of a protein is predicted and then the fold is predicted within that structural class. It has been reported earlier that including structural class prediction method as a post-processing filter improves the fold classification prediction [22]. As shown in this paper this class-wise fold prediction which essentially reduces the pair-wise classifiers to be considered while choosing the most optimum fold gives rise to the highest fold prediction accuracies for all structural classes.

3.2 Hierarchical approach to protein fold prediction

In the hierarchical classification scheme, two-level classification scheme is used where at the first level structural class of a protein is predicted and then at the second level its fold is predicted within that structural class. Since hierarchical approach needed a method for structural class prediction, we used our own state of the art method for protein structural class prediction [23] as the structural class-predictor in the hierarchical approach of fold-recognition. It is a SVM-based method for protein structural class prediction that uses features derived from the predicted secondary structure and predicted burial information of amino acid residues [23]. We would like to mention here that this protein structural class prediction method was developed as a prelude to the development of hierarchical approach to fold prediction. We developed this method based on secondary structure and burial information for easy integration with our fold prediction method [4] and keeping in mind the eventual development of hierarchical approach to fold prediction. Once the structural class is identified the query sequence is subjected to fold-identification within that structural class.

Since in the present study folds belonging to only four major structural classes have been considered, four separate SVMs were trained using optimized parameters for folds of each structural class: (1) for all-α class folds, (2) for all-β class folds, (3) for α/β class folds and (4) α+β class folds; and five-fold cross-validation accuracies were calculated for the class-wise fold predictions. For example in case of 79-Folds dataset; 17 folds belong to all-α structural class, 22 folds to all-β class, 21 folds to α/β class & 19 folds to α+β class (Table 2) and hence four separate SVMs were trained for 17-class, 22-class, 21-class and 19-class classifications. Therefore, use of hierarchical approach led to reduction of a fairly large 79-class classification problem into four much less complicated 17-class, 22-class, 21-class and 19-class classification problems. The class-wise fold prediction accuracies obtained for different structural classes are given in Table 3. Five-fold cross-validation accuracies of 62%, 60%, 55% and 53% were obtained for all-α class folds, all-β class folds, α/β class folds, and α+β class folds, respectively. Class-wise fold prediction helped in improving the overall accuracy of prediction to 57% for 79-Folds dataset, which is significantly better than the previous five-fold cross-validation accuracy of 50%.

Furthermore, we also carried out the similar studies for the 176-Folds dataset. This dataset contains 39 folds belonging to all-α structural class, 38 folds to all-β class, 46 folds to α/β class & 53 folds to α+β class (Table 2). Four separate SVMs were trained for folds belonging to four different structural classes and five-fold cross-validation accuracies were calculated. The class-wise fold prediction accuracies obtained for different structural classes using 176-Folds dataset are given in Table 3. We found that here too hierarchical scheme led to the significant improvement in overall prediction accuracy from 44% to 50%. These results show that fold prediction is improved to a great extent when it is narrowed down to class level owing to reduction in the classification complexity.

Table 3: The fold prediction accuracies (%) obtained when fold prediction is carried out directly and also when fold prediction is done hierarchically at individual structural class-level. The accuracies reported here are five-fold cross-validation accuracies.

Datasets	Direct Fold prediction Accuracy	Hierarchical approach – Class-wise fold predictions				
		All-α Class Folds	All-β Class Folds	α/β Class Folds	α+β Class Folds	Overall prediction accuracy across all classes
79-Folds dataset	49.5	61.5	60.1	55	53.4	**57.3**
176-Folds dataset	43.6	52.2	54.3	47.7	43.6	**49.6**

3.3 Fold prediction using 711F dataset

We extended our studies to the 711F dataset. It may be noted here that 711F dataset consist of proteins belonging to 711 different SCOP folds out of which 187 folds belong to all-α structural class, 130 folds to all-β class, 129 folds to α/β class & 265 folds to α+β class (Table 2). By extending fold prediction studies to this large dataset of 711 folds, we are able to perform fold recognition for 711 folds, which is to a certain extent an approximation of real-world situation. The new method covers about 60% (711/1195) of all known folds and hence probability of any newly sequenced protein to be the prediction target of our method would be around 60% given all the folds had similar likelihood of occurrence. However, it

may be noted here that certain folds do occur more frequently than others and hence the likelihood for any new protein to be the target of this method will be actually much higher.

Fold prediction within structural classes were carried on the 711F dataset by training four separate SVMs. The five-fold cross-validation accuracy, sensitivity and specificity values were computed for fold prediction within each structural class and are given in Table 4. The five-fold prediction accuracies of 84%, 83%, 74% and 84% were obtained for all-α class folds, all-β class folds, α/β class folds, and α+β class folds, respectively. The corresponding sensitivity and specificity values of 79% & 88% was obtained for all-α class folds, 76% & 90% for all-β class folds, 65% & 82% for α/β class folds, and 80% & 92% for α+β class folds, respectively. Our new method of protein fold prediction achieves an overall five-fold cross-validation accuracy of ~80% across all the folds. However, it may be noted here that though fold prediction is improved due to reduction in the classification complexity by the use of hierarchical scheme, contribution to the higher fold-prediction accuracy can also be due to the presence of related sequences in the dataset, especially in a dataset like 711F dataset, which have accumulated because sequences were included to the dataset for folds having less than ten structures, based on sequence similarity. In general, the prediction system gives high accuracy on dataset having related sequences. In order to test influence of such a bias an independent dataset was used in further evaluations.

Table 4: The five-fold cross-validation accuracies, sensitivities and specificities obtained for individual structural-class folds using 711F dataset.

Structural Class	Accuracy (%)	Sensitivity (%)	Specificity (%)
all-α class folds	83.9	79.1	87.9
all-β class folds	82.6	75.9	90.4
α/β class folds	74.3	64.7	81.9
α+β class folds	84.1	80.1	92.3

3.4 Performance evaluation on Lindahl and Elofsson's dataset

The best way to judge the performance of any method is to evaluate its comparative performance on a benchmark dataset. We use the Lindahl and Elofsson's dataset [24] – a large benchmark dataset derived from the SCOP database [20], which has been commonly used for benchmarking of fold recognition methods [1, 2, 8, 9, 25]. This dataset contains 976 proteins and no two proteins in this dataset have >40% pair wise sequence identity to each other. We tested the performance of our method on Lindahl and Elofsson's dataset and compared with benchmarking results reported on this dataset by other taxonomic fold recognition methods like ACCFold [2] and TAXFOLD [1]. Recently, it has been reported that TAXFOLD and ACCFold are the two best performing taxonomy-based fold recognition methods [1]. For the sake of completion, we have also included the prediction accuracies of the template-based fold recognition methods on this dataset as reported in the literature [25]. As evident from the Figure 2, our method gives an accuracy of ~70% whereas TAXFOLD [1] give an accuracy of 41% (Figure 2). The prediction accuracies of other methods are comparatively low. TAXFOLD uses PSI-BLAST profiles and secondary structure information from PSIPRED profiles to discriminate between two folds while our method used features derived from secondary structure information from PSIPRED and burial information from ACCpro. This shows the discriminatory potential of structural-based features for protein fold prediction.

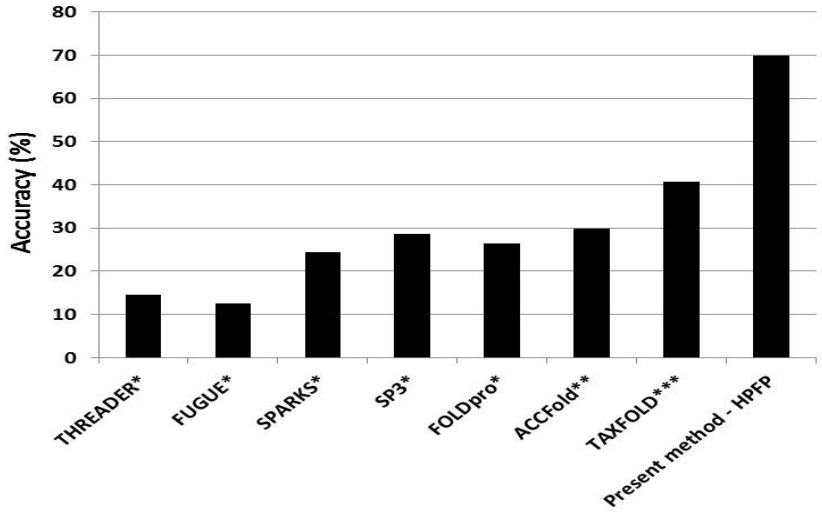

Figure 2: The fold prediction accuracies obtained for different fold prediction methods on Lindahl and Elofsson's dataset. [[1]Only top 1 accuracies reported for all the methods. * The results were cited from Ref 25. **The result was cited from Ref 2. ***The result was cited from Ref 1.]

3.5 Performance evaluation on an independent dataset

We further evaluated the performance of our method on an independent dataset, which contains 2453 protein structures corresponding to 711 different folds. This dataset was derived from the recently released ASTRAL SCOP 1.75. All the newly inducted protein domains in ASTRAL SCOP 1.75 release, which were not there in ASTRAL SCOP 1.73 release, were short listed and similar protein domains were removed from the short listed dataset in such a way that no two proteins have 90% or more than 90% sequence identity to each other. Finally the performance of newly developed method was evaluated on an independent dataset. The evaluation of performance of the new method on this independent data shows that the new method correctly predicts the structural class of 2107 protein sequences and fold of 1490 protein sequences out of 2453 protein sequences. This amounts to the structural class prediction accuracy of (2107/2453) ~ 86% and the fold-prediction accuracy of (1490/2453) ~61%. The respective sensitivity and specificity values for fold prediction are 66% and 76%, respectively. The performance on the independent dataset shows that the new method is working and generalizing well.

Further, the results were analysed to find out where predictions are going wrong. As evident from the results on the independent dataset, a substantial proportion (~40%) of the wrong predictions at the fold level were in fact wrongly predicted at the class level and hence wrong prediction at the fold level is because of wrong prediction at the class level. Therefore, further analysis was done to ascertain whether the poor prediction at the fold level is because of the wrong prediction at the class level or not. All those proteins of the independent dataset, which was predicted wrongly at the both class and fold level were selected and their fold prediction was done in their correct class. It was interesting to observe that around 30% of them were predicted correctly. This shows that the improvement at the class level prediction will further enhance the prediction capability of the new hierarchical method of fold prediction. However, it may be noted here that predictions can go wrong also at the secondary structure and solvent accessibility prediction level as reported earlier [4].

4 Conclusions

In this study we presented our efforts to develop a new method for protein fold prediction with high fold-coverage and high prediction accuracy. The fold-coverage of the method has been increased by increasing the pair-wise sequence identity of the dataset and also by decreasing the cut-off for the minimum number of protein domains per fold. Inclusion of orphan and the folds with fewer than 10 structures led to the increase in total number of folds to 711. The method presented here is the first fold recognition method to include orphan folds to the fold library. Further, the effect of increase in number of folds on prediction performance was tested. It was observed that the prediction accuracy decreases with increase in number of folds. To deal with such large number of folds, hierarchical approach for protein fold recognition was proposed. In the hierarchical approach for protein fold recognition instead of directly predicting the fold of a protein, first structural class of a protein is predicted followed by fold prediction within that structural class. Hence one could reduce the complexity of a large-scale classification problem to a great extent by using hierarchical approach of protein fold recognition. Hierarchical approach leads to the improvement in the overall fold prediction accuracy. Performance evaluation on a benchmark dataset reveals that our method gives state of the art performance. The method presented in this study is the first taxonomic method of protein fold recognition to deal with such large number of folds (711 folds), which is an approximation of real-world situation where number of fold is around 1200. Therefore, the likelihood of any newly sequenced protein to be the prediction target of our method would be about 60% (711/1195). Our new SVM-based method outperforms other available methods and therefore can be used for prediction of most likely fold of proteins discovered in various genome projects. Hence, this can serve as invaluable annotation tool in genome research. Our tool can re-examine proteins hitherto classified as 'unknown' or ‘hypothetical’ and even those annotated as ‘putative’. This method can be made part of any software suite for high throughput protein structure prediction or can be used as stand-alone program.

Acknowledgements

HAN gratefully acknowledges the core funding from CDFD. TASM is thankful to the Council of Scientific and Industrial Research (CSIR) and CDFD for research fellowships. The authors thank colleagues, Rachita and Manjari, for their help. Computational facilities of the SUN Centre of Excellence, CDFD is gratefully acknowledged. Finally, the authors gratefully acknowledge the two anonymous referees for their critical comments.

References

[1] J. Y. Yang and X. Chen. Improving taxonomy-based protein fold recognition by using global and local features. *Proteins,* 79:2053-2064, 2011.

[2] Q. Dong, S. Zhou, and J. Guan. A new taxonomy-based protein fold recognition approach based on autocross-covariance transformation. *Bioinformatics,* 25:2655-2662, 2009.

[3] I. Melvin, E. Ie, R. Kuang, J. Weston, W. N. Stafford, and C. Leslie. SVM-Fold: a tool for discriminative multi-class protein fold and superfamily recognition. *BMC Bioinformatics,* 8 Suppl 4:S2, 2007.

[4] M. T. Shamim, M. Anwaruddin, and H. A. Nagarajaram. Support Vector Machine-based classification of protein folds using the structural properties of amino acid residues and amino acid residue pairs. *Bioinformatics,* 23:3320-3327, 2007.

[5] K. Chen and L. Kurgan. PFRES: protein fold classification by using evolutionary information and predicted secondary structure. *Bioinformatics,* 23:2843-2850, 2007.

[6] H. B. Shen and K. C. Chou. Ensemble classifier for protein fold pattern recognition. *Bioinformatics,* 22:1717-1722, 2006.

[7] L. A. Kelley, R. M. MacCallum, and M. J. Sternberg. Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J Mol Biol,* 299:499-520, 2000.

[8] J. Shi, T. L. Blundell, and K. Mizuguchi. FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *Journal of Molecular Biology,* 310:243-257, 2001.

[9] D. T. Jones, W. R. Taylor, and J. M. Thornton. A new approach to protein fold recognition. *Nature,* 358:86-89, 1992.

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

[11] A. Andreeva, D. Howorth, J. M. Chandonia, S. E. Brenner, T. J. Hubbard, C. Chothia, and A. G. Murzin. Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res,* 36:D419-425, 2008.

[12] P. Deschavanne and P. Tuffery. Enhanced protein fold recognition using a structural alphabet. *Proteins,* 76:129-137, 2009.

[13] K. Kavousi, B. Moshiri, M. Sadeghi, B. N. Araabi, and A. A. Moosavi-Movahedi. A protein fold classifier formed by fusing different modes of pseudo amino acid composition via PSSM. *Computational Biology and Chemistry,* 35:1-9, 2011.

[14] J. M. Chandonia, G. Hon, N. S. Walker, L. Lo Conte, P. Koehl, M. Levitt, and S. E. Brenner. The ASTRAL Compendium in 2004. *Nucleic Acids Res,* 32:D189-192, 2004.

[15] A. Bateman, E. Birney, L. Cerruti, R. Durbin, L. Etwiller, S. R. Eddy, S. Griffiths-Jones, K. L. Howe, M. Marshall, and E. L. Sonnhammer. The Pfam protein families database. *Nucleic Acids Res,* 30:276-280, 2002.

[16] C. Cortes and V. Vapnik. Support-Vector Networks. *Machine Learning,* 20:273-297, 1995.

[17] C. C. Chang and C. J. Lin. LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm. ed, 2001.

[18] L. J. McGuffin, K. Bryson, and D. T. Jones. The PSIPRED protein structure prediction server. *Bioinformatics,* 16:404-405, 2000.

[19] J. Cheng, A. Z. Randall, M. J. Sweredoski, and P. Baldi. SCRATCH: a protein structure and structural feature prediction server. *Nucleic Acids Res,* 33:W72-76, 2005.

[20] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology,* 247:536-540, 1995.

[21] A. L. Cuff, I. Sillitoe, T. Lewis, A. B. Clegg, R. Rentzsch, N. Furnham, M. Pellegrini-Calace, D. Jones, J. Thornton, and C. A. Orengo. Extending CATH: increasing coverage of the protein structure universe and linking structure with function. *Nucleic Acids Research,* 39:D420-426, 2011.

[22] L. Kurgan, K. Cios, and K. Chen. SCPRED: accurate prediction of protein structural class for sequences of twilight-zone similarity with predicting sequences. *BMC Bioinformatics,* 9:226, 2008.

[23] T. A. Mohammad and H. A. Nagarajaram. Svm-based method for protein structural class prediction using secondary structural content and structural information of amino acids. *Journal of Bioinformatics and Computational biology,* 9:489-502, 2011.

[24] E. Lindahl and A. Elofsson. Identification of related proteins on family, superfamily and fold level. *J Mol Biol,* 295:613-625, 2000.

[25] J. Cheng and P. Baldi. A machine learning information retrieval approach to protein fold recognition. *Bioinformatics,* 22:1456-1463, 2006.