Interdiscip Sci Comput Life Sci (2014) 6: 176–186
DOI: 10.1007/s12539-013-0199-0

Automatic Classification of Protein Structures Using Physicochemical Parameters

Abhilash Mohan‡, M. Divya Rao‡, Shruthi Sunderrajan, Gautam Pennathur*
(The Center for Biotechnology, Anna University Chennai-600025, Tamilnadu, India)

Received 3 May 2013 / Revised 12 November 2013 / Accepted 5 December 2013

Abstract: Protein classification is the first step to functional annotation; SCOP and Pfam databases are currently the most relevant protein classification schemes. However, the disproportion in the number of three dimensional (3D) protein structures generated versus their classification into relevant superfamilies/families emphasizes the need for automated classification schemes. Predicting function of novel proteins based on sequence information alone has proven to be a major challenge.

The present study focuses on the use of physicochemical parameters in conjunction with machine learning algorithms (Naive Bayes, Decision Trees, Random Forest and Support Vector Machines) to classify proteins into their respective SCOP superfamily/Pfam family, using sequence derived information. Spectrophores™, a 1D descriptor of the 3D molecular field surrounding a structure was used as a benchmark to compare the performance of the physicochemical parameters. The machine learning algorithms were modified to select features based on information gain for each SCOP superfamily/Pfam family. The effect of combining physicochemical parameters and spectrophores on classification accuracy (CA) was studied.

Machine learning algorithms trained with the physicochemical parameters consistently classified SCOP superfamilies and Pfam families with a classification accuracy above 90%, while spectrophores performed with a CA of around 85%. Feature selection improved classification accuracy for both physicochemical parameters and spectrophores based machine learning algorithms. Combining both attributes resulted in a marginal loss of performance. Physicochemical parameters were able to classify proteins from both schemes with classification accuracy ranging from 90-96%. These results suggest the usefulness of this method in classifying proteins from amino acid sequences.

Key words: protein classification, machine learning algorithms, physicochemical parameters, feature selection, svm, random forest, naïve bayes, decision tree, SCOP classification, Pfam classification.

1 Introduction

Over the last few years there has been an exponential rise in non-redundant protein structures deposited in the protein databank (PDB). Ascribing structure and then function to proteins has been a challenge. The methods typically used for this purpose are homology modeling and threading. While homology modeling depends on the sequence similarity between proteins, threading depends on the ability to find similar structural motifs and uses them to construct a composite structure. Threading methodologies depend on structural and sequence classification databases like SCOP and Pfam (Söding *et al.*, 2005). The SCOP database relies on both manual and automatic curation methods (Murzin *et al.*, 1995). Since the last major SCOP release (1.75), the number of non-redundant structures deposited in PDB has increased by around 20% (Santini *et al.*, 2012). This rapid increase necessitates the development of fast automated approaches to keep the databases updated.

Proteins have been classified based on structural and sequence similarity. Protein structural similarity has been described in various terms; FSSP (families of structurally similar proteins) for instance, uses only alignments (Holm and Sander, 1996). Several other methods have been proposed from shape-based descriptors (Røgen and Fain, 2003) to knot-fitting-inspired methods (Erdmann, 2005) to qualify and quantify structural similarity. Automated protein structural classification methods have relied on machine learning (Jain and Hirst, 2010), clustering and graph theory based approaches (Ashby *et al.*, 2013; Kim and Patel, 2006).

Sequence based protein classification techniques have

*Corresponding author.
E-mail: pgautam@annauniv.edu
‡–Equal contribution

focused on identifying common motifs and domains and using these to cluster them together. The underlying concept used by all these methods is that sequence homology implies 'evolutionary connectedness' and hence most methods rely on sequence similarity for classification (Altschul *et al.*, 1990; Gonnet *et al.*, 1992; Henikoff and Henikoff, 1992; Pearson, 1991).

However, classifying proteins into different structural folds based on amino acid sequences has been a challenge. Structural similarity is often used to annotate protein function and structures in general convey more information than amino acid sequences. Genomic high throughput methods have vastly increased the number of unannotated protein sequences. Determining 3D structure is a time consuming process and methods to *ab-initio* predict protein folds with high accuracy from amino acid sequences will greatly improve functional annotation.

In the present study we extracted physicochemical (PC) parameters from protein sequences and used machine learning algorithms (Naive Bayes, Decision Tree, Random Forest and Support Vector Machines) to predict protein classification at the SCOP superfamily and Pfam family levels. Spectrophores^{TM} (Thijs *et al.*, 2011) was used as a benchmark to compare the performance of the PC parameters in protein classification. Further, we modified the machine learning algorithms used in the study to select the most important features and use them for classification. Finally, we combined both descriptors and evaluated performance. We conclude that the physicochemical parameters, on an average, for both SCOP superfamily and Pfam family classification outperform the spectrophore descriptors and also do better than the combined parameters.

2 Materials and Methods

The basic workflow has been shown in Fig. 1. For this study, protein databank structures from 125 Pfam families (v.26) and 165 SCOP superfamilies (v 1.75) were extracted. The Structure Integration with Function, Taxonomy and Sequence (SIFTS) initiative was used to find the mappings between Pfam families and the corresponding PDB structures. These were deposited in a SQLite database and families with at least one hundred PDB structures were used in the subsequent experiments. The dataset creation was carried out by using PDB files for a specific family and other PDB files were randomly picked from other Pfam families. Spectrophores were extracted for all the families and these were used as the input for the machine learning algorithms, this was treated as the experiment dataset. The same dataset was also used for calculating the physicochemical characteristics as previously described (Mohan *et al.*, 2010). A similar procedure was adopted to pick protein structures from SCOP superfamilies. ORANGE a component based machine learning software package was used in the study (Demsar *et al.*, 2004). The individual machine learning algorithms were also modified to enable feature selection (described later). The algorithms were run with a ten-fold cross validation which ensures that each data point is in the test set at least once. To further validate the performance of the algorithms a 70-30% random split of the data was performed. 70% of the dataset was used for training with ten-fold cross validation (CV) while 30% (unseen during training) was held out for testing.

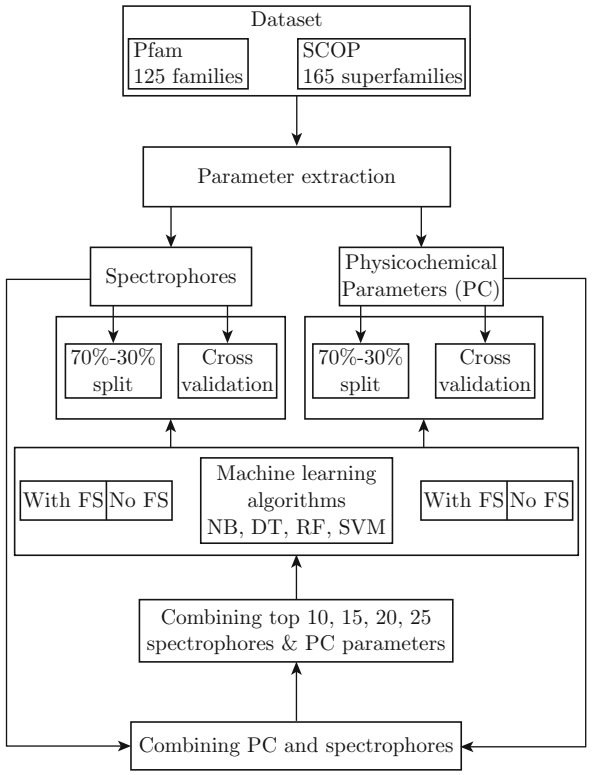

Fig. 1 The workflow of protein classification using Pfam and SCOP datasets. Machine learning algorithms were trained with physicochemical parameters and spectrophores; validation was performed using 70% for training and 30% for testing, ten-fold cross-validation was also performed. Feature selection (FS) based on information gain was implemented and performance between FS and default machine learning algorithms (No FS) was compared. Both descriptors were combined and performance was evaluated, finally the top 10-25 parameters from both feature sets were combined and studied

2.1 Descriptors used in the study

Physicochemical parameters provide information about the protein sequence. They are regarded as being evolutionarily conserved and have been used in identifying protein-protein interactions and sub-cellular localization of proteins (Shen *et al.*, 2007), (Bhasin and Raghava, 2004). Sixty six physicochemical parameters

have been used in this study; they have been tabulated and can be found in the supplementary material. The parameters were extracted using a combination of the "pepstats" program in EMBOSS (Rice *et al.*, 2000) and in-house scripts. Some of parameters used were: number of residues, molecular weight, charge, molar extinction coefficient, isoelectric point, Dayhoff statisitcs and probability of expression in inclusion bodies.

Aliphatic index was calculated using the formula x* (ALA) + a*x (VAL) + b*x (LEU) + b*x (ILE) where a = 2.9 and b = 3.9 are constants. They demonstrate the relative volume of valine and leucine/isoleucine side chains in comparison to the side chains of alanine respectively (Atsushi, 1980). The composition of dipeptides and tripeptides and their frequency of occurence have been associated with solubility and folding of over-expressed proteins (Chan and Dill, 1994). Dipeptide frequency was calculated using $D_{AB,resi} = \log[F_{AB,resi}/F_{AB,nat}]$ where AB is the dipeptide, $F_{AB,resi}$ represents the frequency of dipeptide AB for a specific residue in the protein and $F_{AB,nat}$ is the naturally occurring frequency of the dipeptide. The dipeptide score (SDP) for each protein was calculated using the formula

$$SDP, protein = 1/L - 1 \sum_{resi=1}^{L-1} D_{\mathrm{AB},resi}$$

where L represents the number of dipeptides present in the protein (Idicula-Thomas and Balaji, 2005). The tripeptide score was calculated similarly. Grand average hydropathy (GRAVY) was calculated as the sum of hydropathy values of all the amino acids, divided by the total number of residues present in the protein sequence. The aliphatic index, dipeptide, tripeptide scores and the grand average hydropathy (GRAVY) were calculated using in-house scripts.

Spectrophores are 1D descriptors generated from the property fields of a molecule and are dependent on the actual 3D conformation of the molecule. A spectrophore is calculated by surrounding the 3D structure of a molecule by an artificial "cage" whose interaction with the atom's properties is calculated (Thijs *et al.*, 2011). The interaction thus calculated represents an optimum affinity value between the molecule and the cage. The basic parameters used for calculating a spectrophore are atomic partial charges, atomic lipophilicities, atomic shape deviation and atomic electrophilicities. The atomic partial charges were calculated using the molecular orbital calculations (Momany, 1978) while the atomic electrophilicities have been calculated using Electronegativity Equalization Method (EEM) (Bultinck *et al.*, 2002). The atomic lipophilicities, given by the octanol-water partition coefficient (log P) (Ooms *et al.*, 1998) was calculated using the atomic based approach (Wildman and Crippen, 1999). Atomic shape deviation is the average deviation of each atom from the average molecular radius. Twelve spectrophore values for each property were calculated; these corresponded to the average contribution of the atoms found at the midpoint of each edge of the "cage" surrounding the molecule. A total of forty-eight spectrophore values were generated corresponding to the four parameters utilized.

2.2 Feature Selection

The contribution of every parameter in classification is not uniform; certain parameters contribute more information than others. In this study, we have used information gain, a popular criterion used for feature selection in text categorization (Wang and Lochovsky, 2004) and more recently in image analysis (Dhir *et al.*, 2007). Information gain gives the relative entropy contributed by a specific parameter (Mohan *et al.*, 2010).

2.3 Combining physicochemical and spectrophore parameters

Spectrophores are 3D coordinate descriptors (structure-based) while physicochemical characteristics represent sequence-based data. The combination of these factors may be beneficial. The performance of the algorithms when an equal number of physicochemical and spectrophore parameters were used for the classification of Pfam and SCOP data, was also studied. The number of parameters chosen were varied from 10, 15, 20 and 25 based on the information gain and the effect on classification accuracy was studied.

2.4 Algorithms used in the study

The Naive Bayes (NB) (Hand and Yu, 2001), Decision Tree (DT) (Rasoul and David, 1991), Random Forest (RF) (Livingston, 2005) and Support Vector Machine (SVM) algorithms were used in the study (Vasanthanathan *et al.*, 2009).

The Naive Bayes classifier was set to use "relative frequency" for the estimation of prior probabilities. The LOESS (locally weighted scatter plot smoothing) (Frank *et al.*, 2002) smoothing of the classifier was enabled. The size of the LOESS window set at 0.5 and 100 points were set for interpolating the curve. A trial and error approach was used to optimize these values. The attribute selection criterion was set to "information gain" for the decision trees algorithm and binarization was disabled. Pre-pruning of minimum instance in leaves was left at the default setting of 2 and post pruning was performed using the m-estimate with 'm' being set to 2 and the leaves were recursively merged with the same majority class (Esposito *et al.*, 1997). For the random forest algorithm, the number of trees in the forest was set to 50 and nodes with 5 or fewer instances were stopped from getting split. The C-type of SVM was used with RBF as the kernel function. A grid search

to pick the optimal cost (C) and γ using 10-fold cross validation was performed. The optimal values for the present analysis was C = 4 and $\gamma = 0.015$. Since the C, γ values after the grid search analysis were more or less similar for both spectrophores and physicochemical based feature set; these values were retained for both.

2.5 Modified machine learning algorithms

The ideal set of features may be different for different SCOP superfamilies/Pfam families, in order to pick the best parameters, the machine learning algorithms were modified so as to select five parameters that contributed maximum information gain for each superfamily/family. The four algorithms, NB, DT, RF and SVM were modified accordingly and this experiment was performed for physicochemical and spectrophore parameters for the Pfam and SCOP datasets.

2.6 Performance metrics used in the study

Classification accuracy (CA) estimates the percentage of accurate predictions for a particular object with respect to a specific class (Ankerst *et al.*, 1999). Some of the other metrics used to assess the performance of the classifiers were sensitivity, specificity, the Brier score, area under the receiver operator characteristic (ROC) Curve (AUC), Information score (IS), Matthews Correlation Coefficient (MCC), F1/F2 statistics and Scotts Pi scores. The complete description of these parameters has been included in the supplementary material.

3 Results

The aim of this study was to evaluate physicochemical parameters in conjunction with machine learning algorithms in protein classification. We have used spectrophores, a one-dimensional descriptor of the three-dimensional molecular field surrounding a protein structure (Thijs *et al.*, 2011) in conjunction with the machine learning algorithms as a benchmark to compare the performance of physicochemical parameters.

Four different sets of experiments were performed with two datasets, the SCOP superfamily and Pfam family. The performance of the different machine learning algorithms, NB, DT, RF and SVM, for the two datasets was evaluated using a ten-fold cross validation (CV) and further validated by splitting the dataset (70%-30%). The performance of the classifiers was improved by identifying the top five features contributing to maximum information gain and using only these for classification. The third experiment was performed by combining the physicochemical and spectrophore derived parameters and the effect on classification accuracy was studied for both datasets. Finally, the top 10, 15, 20 and 25 parameters from each set of descriptors (selected based on their information gain) was combined and studied.

The machine learning algorithms were run on the datasets and different statistical parameters like classification accuracy, area under the ROC curve, sensitivity and specificity, Matthews Correlation Coefficient etc. were used to judge the performance of the algorithms. For the sake of brevity, only classification accuracy has been considered for discussion while the complete set of results along with the various metrics used to judge the performance of the algorithm has been tabulated in the supplementary material (Table S1.1).

In order to validate the algorithms, the dataset was randomly split into 70% training set and a separate 30% test set (70-30 split). The algorithms were trained using 70% of the dataset (using cross validation) and the remaining 30% of the data, unseen by the algorithms were used for testing. The experiments were performed using physicochemical parameters and spectrophores descriptors (Fig. 2).

The classification accuracy ranged from 0.92-0.95 and 0.85-0.90 for the Pfam dataset using PC and spectrophores. For the SCOP superfamilies the results were similar with CA ranging between 0.90- 0.94 and 0.84-0.87 respectively. To assess the importance of feature selection we performed a similar experiment with feature selection enabled descriptors i.e. using only the top five attributes with maximum information gain. The results indicated a marginal improvement of classification for all classifiers except Naive Bayes. The CA ranged from 0.89-0.96 and 0.83-0.90 for the Pfam dataset, the results for the SCOP dataset was 0.87-0.96 and 0.80-0.88.

The learners trained with PC parameters consistently performed with classification accuracies above 90% indicating the reliability of this method in classifying proteins into different folds using only sequence information.

A ten-fold cross validation was performed on all the datasets for all the classifiiers. Classification accuracies of 0.92-0.96 and 0.86-0.91 were observed for the Pfam families using both PC and spectrophores (Figs. 7, 8). Similar results were observed for the SCOP dataset with values ranging from 0.91-0.94 and 0.82-0.88 respectively (Figs. 5, 6). Feature selection was performed to improve the performance of the algorithms and these results have been discussed later in detail. Since, there was no significant difference in classification accuracy between the two methods (Supplement 3 (Tables S3. 1, 2, 3, and 4)), henceforth only results achieved using the ten-fold cross validation experiments will be discussed.

3.1 Comparison of physicochemical and spectrophore parameters

The physicochemical parameters classified both the SCOP and Pfam dataset more accurately than spectrophores (Figs. 3, 4), which is surprising since, spectrophores, being a descriptor for 3D coordinates was ex-

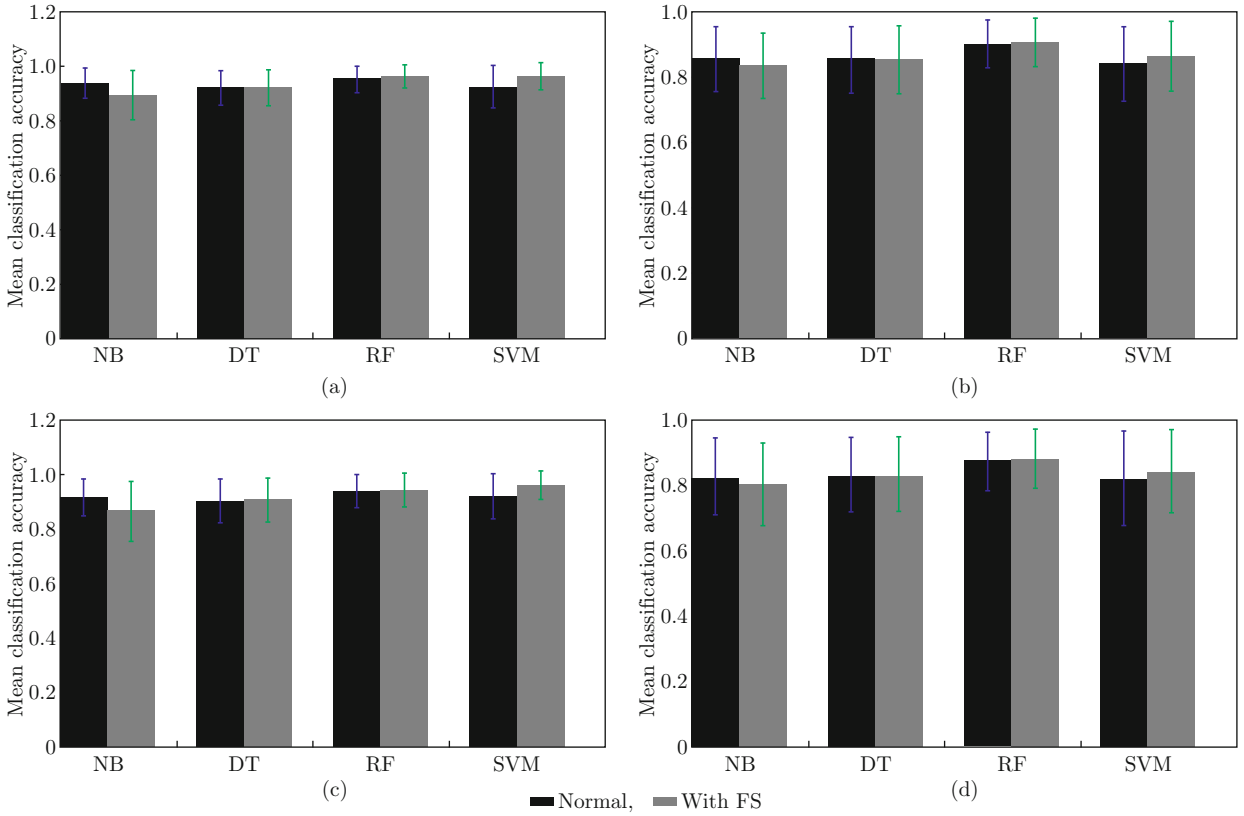

Fig. 2 Comparison of classification accuracies for both the Pfam ((a), (b)) and SCOP ((c), (d)) datasets using physicochemical parameters and Spectrophores. The error bars indicate the standard deviation of the mean for the 125 Pfam families and 165 SCOP superfamilies

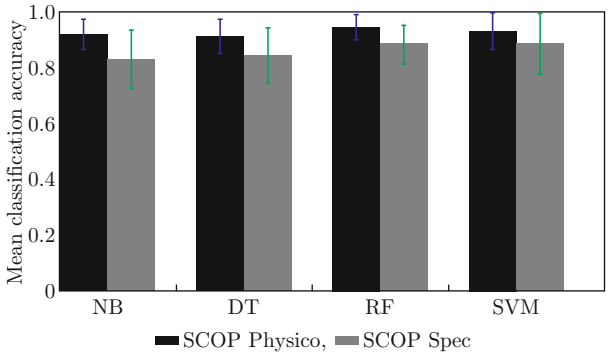

Fig. 3 Comparison of the classification accuracy of SCOP superfamilies using physicochemical and spectrophore parameters. The error bars indicate the standard deviation of the mean for the 165 SCOP superfamilies

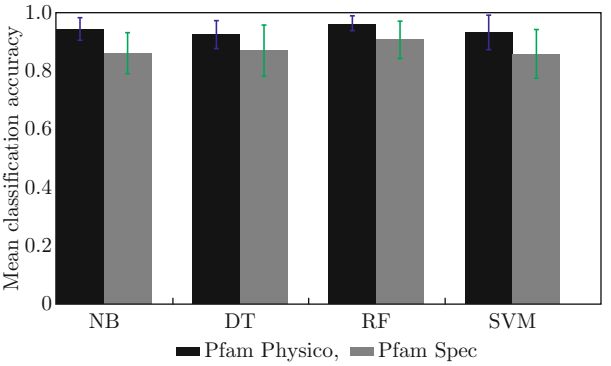

Fig. 4 Comparison of the classification accuracy of Pfam families using physicochemical and spectrophore parameters. The error bars indicate the standard deviation of the mean for the 125 Pfam families

pected to perform better than sequence based descriptors especially for the SCOP dataset. With respect to the Pfam dataset, a 10% improvement in classification accuracy was seen for certain algorithms when compared with spectrophore descriptors, while a 11% increase in classification accuracy was seen in NB classifier for the SCOP dataset. The algorithms trained with PC parameters performed better than the spectrophore descriptors for both datasets.

3.2 Comparison of native and feature selection enabled algorithms

It is well known that machine learning algorithms when faced with redundant features tend to be inefficient. Feature selection helps improve accuracy by re-

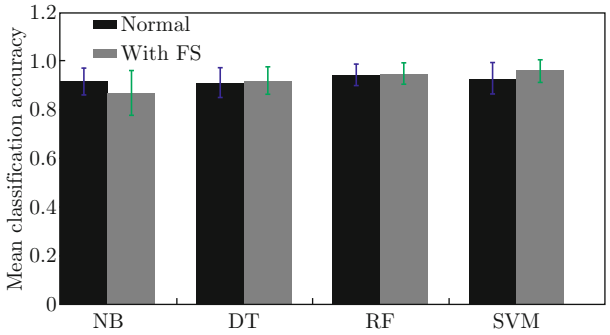

Fig. 5 Comparison of the classification accuracy of SCOP superfamilies using physicochemical parameters with different default machine learning algorithms(normal) and algorithms after feature selection (with FS). The error bars indicate the standard deviation of the mean for the 165 SCOP superfamilies

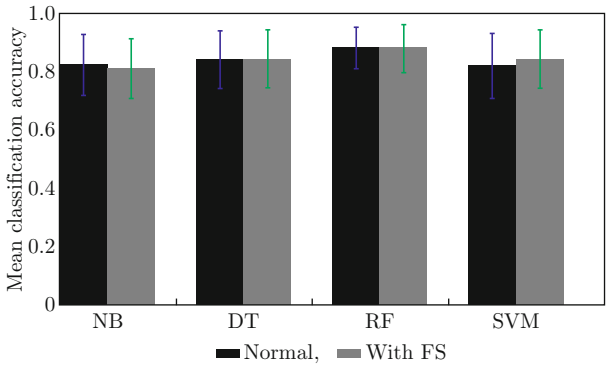

Fig. 6 Comparison of the classification accuracy of SCOP superfamilies using spectrophore descriptors with different default machine learning algorithms(normal) and algorithms after feature selection (with FS). The error bars indicate the standard deviation of the mean for the 165 SCOP superfamilies

moving redundant attributes or those that contribute the least information gain (Lu *et al.*, 2004). The classification accuracy of the SCOP dataset was studied with both the native and feature selection enabled classifiers (trained using PC parameters). Feature selection was found to improve performance in all the learners except NB, SVM showed significant improvement in performance (Fig. 5). A similar trend was observed in the case of spectrophores with SVM showing maximum improvement (Fig. 6). The classification accuracy for individual superfamilies is shown in the supplementary material (Figs. S1-S4 (physicochemical parameters) and Figs. S5-S8 (spectrophore parameters)). The performance of the Pfam data was remarkably similar to the SCOP dataset, a result that was not expected, since Pfam data is based on sequence conservation but SCOP data not only relies on sequence but also on structural domain conservation. SVM showed the maximum improvement with feature selection for both the spectrophore and physicochemical parameters. The classification accuracy for individual Pfam families is shown in the supplementary material (Figs. S9-S12 (physicochemical parameters) and Figs. S13-S16 (spectrophore parameters)). The results for the Pfam dataset are shown in (Figs. 7, 8).

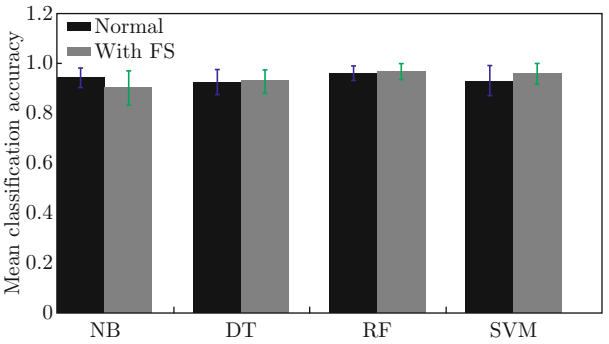

Fig. 7 Comparison of classification accuracy for the Pfam families using physicochemical descriptors with different default machine learning algorithms(normal) and algorithms after feature selection (with FS). The error bars indicate the standard deviation of the mean for the 125 Pfam families

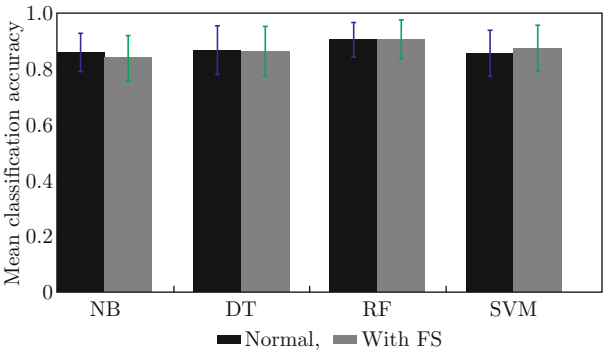

Fig. 8 Comparison of the classification accuracy of Pfam families using spectrophore descriptors with different default machine learning algorithms(normal) and algorithms after feature selection (with FS). The error bars indicate the standard deviation of the mean for the 125 Pfam families

3.3 Combining spectrophore and physicochemical parameters

Spectrophores contribute a very different set of attributes when compared to PC parameters, largely because they represent the molecular field that envelops each molecule and may be considered as spatial attributes. Since these two feature sets are so different from one another they may complement each other and improve performance. A total of 48 spectrophore and 66 physicochemical parameters were combined and the machine learning algorithms were run using the

combined feature set for both datasets. The algorithms trained with PC descriptors outperformed spectrophores and the combined feature set. The average classification accuracy across all machine learning algorithms for spectrophore based classification was 0.843, for the combined it was 0.890 while the physicochemical parameter based classification showed an average of 0.924.

None of the individual algorithms were able to outperform the physicochemical parameter based classification, with the random forest algorithm performing the best among all the classifiers. The random forest algorithm for the combined parameter (CA: 0.914) set came closest to the physicochemical based parameter set (CA: 0.943) (Fig. 9). The classification accuracy for individual SCOP superfamilies is shown in the supplementary material (Figs. S21-S24). The Pfam dataset showed a similar trend to the SCOP dataset with physicochemical parameters outperforming the combined and spectrophore parameter sets. The average classification accuracy across all the machine learning algorithms were 0.873, 0.938 and 0.927 for spectrophores, physicochemical and combined set respectively (Fig. 10). The classification accuracy for individual SCOP superfamilies is shown in the supplementary material (Figs. S17-S20).

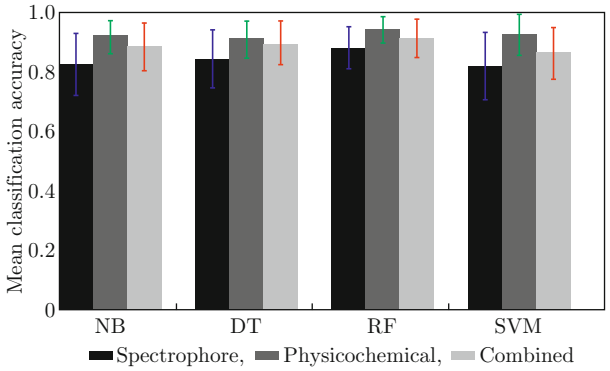

Fig. 9 Comparison of the classification accuracy of SCOP superfamilies using a combination of physicochemical and spectrophore parameters. The error bars indicate the standard deviation of the mean for the 165 SCOP superfamilies

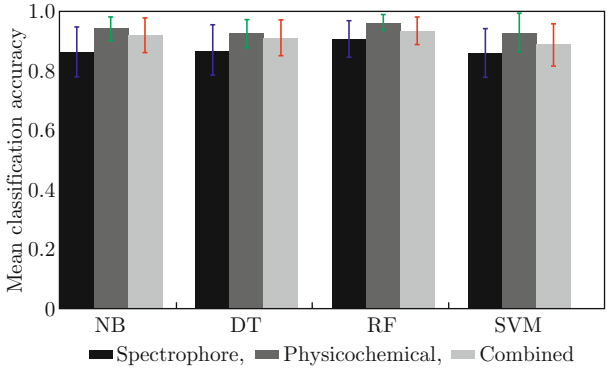

Fig. 10 Comparison of the classification accuracy of Pfam families using a combination of physicochemical and spectrophore parameters. The error bars indicate the standard deviation of the mean for the 125 Pfam families

3.4 Combining top performing spectrophores and physicochemical parameters

It was observed from feature selection experiments that including only top performing features could increase the overall classification accuracy. The top 10, 15, 20, 25 parameters that contributed the maximum information gain for each feature set were combined. The number of combined parameters was restricted to 25 as beyond this there was no performance benefit (data not shown). The classification accuracy of the SCOP dataset using the top 25 spectrophore and physicochemical features was similar to the results obtained when physicochemical parameters were solely used. The random forest algorithm performed the best among the classifiers with the combined parameter set increasing the classification accuracy of certain superfamilies like 51182 (3%), 58119 (3%), 57667 (5%).

The data for individual superfamilies is shown in the supplementary materials (top 10, 15, 20, 25 combinations of parameters correspond to Figs. (S25-S28), (S29-S32), (S33-S36) and (S37-S40) respectively). By implementing feature selection and combining both attribute sets there was a marked improvement in performance. Further, the performance was comparable to the use of physicochemical parameters (Table 2). For instance, the performance was markedly higher in SVM when compared to physicochemical parameters, for the families PF02777 (15% higher), PF00413 (11% higher), PF00127 (6% higher) and PF02777 (5% higher), when the top 25 spectrophore and physicochemical parameters were used. In the case of DT, the combined dataset (top 25 parameters) was found to better the algorithms trained with PC parameters. Individual graphs for the performance of each family is shown in the supplementary materials (top 10, 15, 20, 25 combinations of parameters correspond to Figs. (S41-S44), (S45-S48), (S49-S52) and (S53-S56) respectively).

We were also interested in looking at the impact of combined feature selection on performance, especially of the underperforming Pfam families (PF00149, PF00169, and PF00583). There was a significant improvement, as much as ten percent, in the classification accuracy of these three families in SVM classifier. Random forest also had an improvement however it was not as dramatic as SVM. In other classifiers feature selection improved the CA of some families (PF00169) but not others. With respect to the SCOP

Table 1 The classification accuracy of the combined parameters (10, 15, 20, 25) for the SCOP dataset when compared to individual descriptors

Parameters	Machine Learning Algorithms											
	NB			DT			RF			SVM		
	Spec	Phy	Comb	Spec	Phy	Comb	Spec	Phy	Comb	Spec	Phy	Comb
10			0.897			0.902			0.923			0.885
15	0.826	0.918	0.901	0.842	0.911	0.908	0.882	0.961	0.937	0.822	0.926	0.895
20			0.904			0.911			0.939			0.899
25			0.906			0.913			0.940			0.903

Table 2 The classification accuracy of the combined parameters (10, 15, 20, 25) for the Pfam dataset when compared to individual descriptors

Parameters	Machine Learning Algorithms											
	NB			DT			RF			SVM		
	Spec	Phy	Com	Spec	Phy	Com	Spec	Phy	Com	Spec	Phy	Com
10			0.927			0.917			0.944			0.905
15	0.863	0.941	0.933	0.867	0.923	0.922	0.907	0.961	0.951	0.857	0.928	0.915
20			0.935			0.930			0.954			0.918
25			0.936			0.922			0.955			0.925

dataset, there was an overall improvement in all the classifiers.

4 Discussion

4.1 Pfam families with poor classification accuracy

A classification accuracy of less than 85% was used as a cut-off to identify mediocre performing Pfam families and SCOP superfamilies. The families that fit this criterion among the Pfam dataset are PF00149, PF00169, PF00583. It was observed that several of the poorly performing families were diverse proteins, for instance, PF00149 (metallophos) encompasses a family of calcineurin-like phosphoesterases and these proteins are involved in the regulation of cellular function, including activation or inhibition of enzymes.

It has been previously reported that proteins in this family are involved in metal chelation and this makes classification dificult as metal chelation requires unique binding and catalytic sites (Wu *et al.*, 2003). PF00169, Pleckstrin homology domain is another family that performed badly; they are involved in signalling and are constituents of the cytoskeleton. Variability in the length of the loops linked to the β strands and low sequence similarity may have contributed to its inferior performance (Blomberg and Nilges, 1997).

Finally, PF00583 comprises acetyltransferases belonging to the GNAT family; these proteins are ubiquitous and use acyl-CoA to acylate their cognate substrates. The proteins in this family show extensive sequence divergence (Dyda *et al.*, 2000), and this could be the reason for their poor performance.

4.2 SCOP superfamilies with poor classification accuracy

The outliers in the SCOP dataset were analyzed in a process similar to the Pfam dataset, the SCOP superfamily 54211 under-performed across all the classifiers. Three SCOP superfamilies (50249, 50692 and 50729) performed with a classification accuracy below our cut-off (<85%) in all the classifiers other than SVM. 54211 corresponds to the "ribosomal protein S5 domain 2-like" superfamily with α and β proteins.

It is a large sequence diverse superfamily; studies have found correlations between functional diversity and size of the superfamily and this may explain inaccurate classification (Casbon and Saqi, 2006). 50729 represent the PH domain-like superfamily and consist of β proteins. Interestingly, this superfamily contains the pleckstrin homology domain family i.e. PF00169 and similar reasons may account for its bad performance. The ADC-like superfamily (50692) is considered to be one of a number of superfamilies that change their architecture and topology thereby resulting in a change of their core structure. This family of proteins is found among bacteria, fungi and plants. They catalyze the conversion of L-aspartate to β-alanine and are responsible for β - alanine production which is essential for the biosynthesis of Vitamin B5 (pantothenate). This

superfamily is represented by a unique double-psi β-barrel fold with the characteristic six-stranded β-barrel (Arumugam *et al.*, 2013).

We looked for correlations among the mediocre performers in SCOP and Pfam. Remarkably, two of the SCOP superfamilies (54211 and 50729) corresponded to under-performing Pfam families (PF02518, PF00071 and PF00169). PF00169 is one of the Pfam families that performed badly in all classifiers while the other two protein families under-performed in some of the classifiers but not in all of them.

4.3 The important physicochemical and Spectrophore features identified using modified machine learning algorithms

Since each Pfam family was trained with the top five physicochemical parameters based on their contribution to information gain we analyzed whether some of the parameters were more important than the others. The five best parameters that contributed most to information gain were dynamic and changed with each family. We have identified and tabulated the top ten parameters from this dynamic set (Table 3). The number of amino acid residues was found to be the most important parameter and was used to train thirty-three Pfam families. Protein size is usually represented by the number of amino acids and is an important factor in determining structure-function relationships. The residues are exposed to different "average" environments in different subsets of proteins based on their size. There is a negative correlation between the number of hydrophilic amino acids and protein size, aspartic acid is the only exception with its frequency increasing with increase in size (Shirota *et al.*, 2008). As expected molecular weight is another conserved property and may be a consequence of selecting families of similar lengths (Hobohm and Sander, 1995).

Table 3 Top ten physicochemical parameters that influence classification of protein families (Pfam)

No.	Physicochemical parameters	No. of Pfam Families
1.	Number of residues	33
2.	Probability of expression in the inclusion bodies	32
3.	Molecular weight	31
4.	Isoelectric point	27
5.	Molar extinction coefficient (A280)	23
6.	Dipeptide	21
7.	Dayhoff statistic of methionine	19
8.	Dayhoff statistic of tyrosine	18
9.	Dayhoff statistic of tryptophan	17
10.	% of methionine	17

Dayhoff statistics of tyrosine and tryptophan (aromatic residues) were found to be important descriptors. Aromatic residues are highly conserved in a number of protein families, including membrane proteins where they flank a belt of hydrophobic residues in the lipid facing surface (Elofsson and Heijne, 2007).

In the case of SCOP superfamilies a similar process was applied and the results are show in in Table 4. Many of the parameters that were important in Pfam classification were represented here as well. Dipeptides and tripeptides were important parameters in the classification of SCOP superfamilies, this is understandable because protein folding is dependent on the neighbouring amino acids as well as individual amino acids (Sun and Huang, 2006). Tryptophan, tyrosine and phenylalanine residues are highly conserved and implicated as likely binding site residues, by conferring structural rigidity at binding interfaces and reducing the entropic cost of binding (Ma *et al.*, 2003).

Table 4 Top ten physicochemical parameters that influence classification of protein families (SCOP)

No.	Physicochemical parameters	No. of SCOP Families
1.	Number of residues	57
2.	Molecular weight	50
3.	Dipeptide	44
4.	Isoelectric point	39
5.	Probability of expression in inclusion bodies	31
6.	Molar extinction coefficient	30
7.	Tripeptide	26
8.	% of tyrosine	20
9.	Dayhoff statistic of tyrosine	19
10.	Dayhoff statistic of Phenylalanine	17

5 Conclusion

With the advent of high throughput sequencing, there is a huge disparity between the speed of sequence data production and analysis of this data. Protein classification is the first step to functional annotation and accurate protein classification without any structural information is still a challenge. A number of previous studies have attempted to identify meaningful and accurate features that facilitate protein function prediction using biological approaches and computational tools.

In the present study, we have come up with a set of features that can accurately classify proteins into different SCOP superfamilies/Pfam families using only the amino acid sequence information. The physicochemical parameters based machine learning classifiers per-

formed better in both SCOP and Pfam datasets when compared to spectrophores. Feature selection improved the performance of both the physicochemical & spectrophore based classifiers. But combining the two parameters did not improve classification accuracy.

The main contribution of this work is the identification of a set of physicochemical parameters that can be extracted from the amino acid sequence, and can then be used in conjunction with machine learning algorithms to accurately classify proteins into SCOP superfamilies or Pfam families.

Acknowledgement

PG wishes to acknowledge The Department of Biotechnology (DBT), Govt of India, New Delhi, through the Centre of Excellence (No. BT/01/COE/07/01) and Biotechnology Information System, New Delhi, Govt. of India for financial support. AM would like to thank DBT, New Delhi for fellowship (BT/01/COE/07/01). DR would like to thank The Department of Science and Technology, New Delhi for fellowship through the project SR/SO/BB-28/2008 and SR would like to thank DBT, New Delhi through the project BT/PR11310/PDB/26.160/2008 for fellowship.

Disclosure statement

No competing financial interests exist.

References

[1] Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J. 1990. Basic local alignment search tool. J Mol Biol 215, 403-410.

[2] Ankerst, M., Kastenmüller, G., Kriegel, H.P., Seidl, T., *et al.*, 1999. Nearest neighbor classification in 3d protein databases. In Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology, 34-43.

[3] Arumugam, G., Nair, A.G., Hariharaputran, S., Ramanathan, S. 2013. Rebelling for a reason: Protein structural outliers. PloS one 8, e74416.

[4] Ashby, C., Johnson, D., Walker, K., Kanj, I.A., Xia, G., Huang, X. 2013. New enumeration algorithm for protein structure comparison and classification. BMC Genomics 14, S1.

[5] Atsushi, I. 1980. Thermostability and aliphatic index of globular proteins. J Biochem 88, 1895-1898.

[6] Bhasin, M., Raghava, G. 2004. Eslpred: Svm-based method for subcellular localization of eukaryotic proteins using dipeptide composition and psi-blast. Nucleic Acids Res 32, W414-W419.

[7] Blomberg, N., Nilges, M. 1997. Functional diversity of ph domains: an exhaustive modelling study. Fold Des 2, 343-355.

[8] Bultinck, P., Langenaeker, W., Lahorte, P., De Proft, F., Geerlings, P., Waroquier, M., Tollenaere, J. 2002. The electronegativity equalization method I: Parametrization and validation for atomic charge calculations. J Phys Chem A 106, 7887-7894.

[9] Casbon, J., Saqi, M. 2006. Functional diversity within proteins superfamilies. Journal of Integrative Bioinformatics 3.

[10] Chan, H.S., Dill, K.A. 1994. Transition states and folding dynamics of proteins and heteropolymers. J Chem Phys 100, 9238.

[11] Demšar, J., Zupan, B., Leban, G., Curk, T. 2004. Orange: From experimental machine learning to interactive data mining. Springer, Berlin, Heidelberg, pp 537-539.

[12] Dhir, C., Iqbal, N., Lee, S.Y. 2007. Efficient feature selection based on information gain criterion for face recognition. In Information Acquisition, 2007. ICIA'07. International Conference on. IEEE, 523-527.

[13] Dyda, F., Klein, D.C., Hickman, A.B. 2000. Gcn5-related n-acetyltransferases: a structural overview. Annu Rev Bioph Biom 29, 81-103.

[14] Elofsson, A., Heijne, G.V. 2007. Membrane protein structure: prediction versus reality. Annu Rev Biochem 76, 125-140.

[15] Erdmann, M.A. 2005. Protein similarity from knot theory: geometric convolution and line weavings. J Comput Biol 12, 609-637.

[16] Esposito, F., Malerba, D., Semeraro, G., Kay, J. 1997. A comparative analysis of methods for pruning decision trees. IEEE T Pattern Anal 19, 476-491.

[17] Frank, E., Hall, M., Pfahringer, B. 2002. Locally weighted naive bayes. In Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann Publishers Inc., 249-256.

[18] Gonnet, G.H., Cohen, M.A., Benner, S.A. 1992. Exhaustive matching of the entire protein sequence database. Science 256, 1443-1445.

[19] Hand, D.J., Yu, K. 2001. Idiot's bayes not so stupid after all? Int Stat Rev 69, 385-398.

[20] Henikoff, S., Henikoff, J.G. 1992. Amino acid substitution matrices from protein blocks. P Natl Acad Sci USA 89, 10915-10919.

[21] Hobohm, U., Sander, C. 1995. A sequence property approach to searching protein databases. J Mol Biol 251, 390-399.

[22] Holm, L., Sander, C. 1996. The fssp database: fold classification based on structure-structure alignment of proteins. Nucleic Acids Res 24, 206-209.

[23] Idicula-Thomas, S., Balaji, P.V. 2005. Understanding the relationship between the primary structure of proteins and its propensity to be soluble on overexpression in escherichia coli. Protein Sci 14, 582-592.

[24] Jain, P., Hirst, J.D. 2010. Automatic structure classification of small proteins using random forest. BMC bioinformatics 11, 364.

[25] Kim, Y.J., Patel, J.M. 2006. A framework for protein structure classification and identification of novel protein structures. BMC bioinformatics 7, 456.

[26] Livingston, F. 2005. Implementation of breiman's random forest machine learning algorithm. ECE591Q Machine Learning Journal Paper.

[27] Lu, Z., Szafron, D., Greiner, R., Lu, P., Wishart, D.S., Poulin, B., Anvik, J., Macdonell, C., Eisner, R. 2004. Predicting subcellular localization of proteins using machine-learned classifiers. Bioinformatics 20, 547-556.

[28] Ma, B., Elkayam, T., Wolfson, H., Nussinov, R. 2003. Protein-protein interactions: Structurally conserved residues distinguish between binding sites and exposed protein surfaces. P Natl Acad Sci USA 100, 5772-5777.

[29] Mohan, A., Anishetty, S., Gautam, P. 2010. Global metal-ion binding protein fingerprint: A method to identify motif-less metal-ion binding proteins. J Bioinform Comput Biol 8, 717-726.

[30] Momany, F. 1978. Determination of partial atomic charges from ab initio molecular electrostatic potentials. Application to formamide, methanol, and formic acid. J Phys Chem 82, 592-601.

[31] Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C. 1995. Scop: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247, 536-540.

[32] Ooms, F., Wouters, J., Collin, S., Durant, F., Jegham, S., George, P. 1998. Molecular lipophilicity potential by clip, a reliable tool for the description of the 3d distribution of lipophilicity: application to 3-phenyloxazolidin-2-one, a prototype series of reversible maoa inhibitors. Bioorg Med Chem Lett 8, 1425-1430.

[33] Pearson, W.R. 1991. Searching protein sequence libraries: comparison of the sensitivity and selectivity of the smith-waterman and fasta algorithms. Genomics 11, 635-650.

[34] Rasoul, S., David, L. 1991. A survey of decision tree classifier methodology. IEEE Trans Syst Man Cybern 21, 660-674.

[35] Rice, P., Longden, I., Bleasby, A. 2000. Emboss: the european molecular biology open software suite. Trends Genet 16, 276-277.

[36] Røgen, P., Fain, B. 2003. Automatic classification of protein structure by using gauss integrals. P Natl Acad Sci USA 100, 119-124.

[37] Santini, G., Soldano, H., Pothier, J. 2012. Automatic classification of protein structures relying on similarities between alignments. BMC bioinformatics 13, 233.

[38] Shen, J., Zhang, J., Luo, X., Zhu, W., Yu, K., Chen, K., Li, Y., Jiang, H. 2007. Predicting protein- protein interactions based only on sequences information. P Natl Acad Sci USA 104, 4337-4341.

[39] Shirota, M., Ishida, T., Kinoshita, K. 2008. Effects of surface-to-volume ratio of proteins on hydrophilic residues: Decrease in occurrence and increase in buried fraction. Protein Sci 17, 1596-1602.

[40] Söding, J., Biegert, A., Lupas, A.N. 2005. The hhpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res 33, W244-W248.

[41] Sun, X.D., Huang, R.B. 2006. Prediction of protein structural classes using support vector machines. Amino Acids 30, 469-475.

[42] Thijs, G., Langenaeker, W., De Winter, H. 2011. Application of spectrophores to map vendor chemical space using self-organising maps. J Cheminformatics 3, 1-1.

[43] Vasanthanathan, P., Taboureau, O., Oostenbrink, C., Vermeulen, N.P.E., Olsen, L., Jrgensen, F.S. 2009. Classification of cytochrome p450 1a2 inhibitors and noninhibitors by machine learning techniques. Drug Metab Dispos 37, 658-664.

[44] Wang, G., Lochovsky, F.H. 2004. Feature selection with conditional mutual information maximin in text categorization. In Proceedings of the thirteenth ACM international conference on Information and knowledge management. ACM, 342-349.

[45] Wildman, S.A., Crippen, G.M. 1999. Prediction of physicochemical parameters by atomic contributions. J Chem Inf Comp Sci 39, 868-873.

[46] Wu, C.H., Huang, H., Yeh, L.S.L., Barker, W.C. 2003. Protein family classification and functional annotation. Comput Biol Chem 27, 37-47.