J Mol Model (2012) 18:2741–2754
DOI 10.1007/s00894-011-1223-0

ORIGINAL PAPER

Dimensionality reduction in computational demarcation of protein tertiary structures

Rajani R. Joshi · Priyabrata R. Panigrahi · Reshma N. Patil

Received: 9 June 2011 / Accepted: 11 August 2011 / Published online: 25 November 2011

Abstract Predictive classification of major structural families and fold types of proteins is investigated deploying logistic regression. Only five to seven dimensional quantitative feature vector representations of tertiary structures are found adequate. Results for benchmark sample of non-homologous proteins from SCOP database are presented. Importance of this work as compared to homology modeling and best-known quantitative approaches is highlighted.

Keywords Logistic regression · Principal component analysis · Protein structural classes · Quantitative features of tertiary folds · SCOP database

Introduction

Structural classification of proteins helps deciphering their evolutionary connections and local and tertiary fold relationship between them. Several databases in public domain exist which perform this classification at various hierarchical levels, with different objectives. Principal among them are the SCOP [1] and CATH [2] databases.

R. R. Joshi (✉)
Department of Mathematics,
Indian Institute of Technology Bombay,
Powai,
Mumbai 400076, India
e-mail: rrj@iitb.ac.in

P. R. Panigrahi
Biochem Sci Group, National Chemical Laboratory,
Pune, India

R. N. Patil
Department of Chem. Engg, Indian Institute of Technology Bombay,
Mumbai, India

Computational modeling of protein structures using quantitative data structures offers efficient, cost-effective applications for classification as well as characterization of protein structures, analysis of protein structure-function correlations and understanding of protein structural genomics. Quantitative data structures found computationally feasible in wide-ranging applications of this kind mostly consists of feature vectors, trees, and graphs. While tree or graphs are of direct applications in homology mapping and/or computer aided analysis of molecular recognition, protein binding and functional interactions (e.g., [3–6]), computing with these is more complex and often requires special data mining algorithms and tools as compared to feature vector representation.

Quantitative feature vectors are computationally the simplest data structures. These are also most suitable for applications of theoretically sound statistical data mining techniques. Representation of fixed size segments of protein sequences as quantitative feature vectors has been useful in phylogenic classification and secondary structure analysis of proteins and has also offered applications in ab initio prediction of tertiary structure [7–11].

Chi et al. [12, 13] have used 25-dimensional feature vector for fast protein structure retrieval and fold classification. We have attempted structural classification at the first level of the hierarchy in SCOP considering the local and global quantitative features used by them. Sequential as well as structural similarity is important in homology modeling. In view of this, we considered also incorporating some sequential features which are not a linear combination of the features used by Chi et al., yet which are of the same 'type' in the sense that it pertains to geometry and does not explicitly require the knowledge of which amino acids are there in the sequence and in what order, etc.

Length of a protein sequence is simplest if its linear geometrical features satisfying the above criterion. Our earlier

studies on ab initio prediction of protein tertiary structure using only the primary sequence have shown this feature as a statistically significant variable in correlation of the inter-residue distances in primary and tertiary structures of proteins [10, 14]. Moreover, inclusion of this feature does not increase the complexity of computing the feature-vector, so we have included it along with the features used by Chi et al [12, 13].

Principal component analysis is carried out to get descriptors of these features collectively in a reduced dimensional space. *Multi-class logistic regression* is then applied to provide possible application for predictive classification. The results show significance of specific features in characterizing specific structural families of proteins, and also in identifying different types of folds within a family (class).

Materials and methods: quantitative feature vector representation and analysis

We represent a structured protein as a data point in a 26-dimensional feature space. These 26 features are listed in Table 1. Length of protein sequence, listed as the first global feature in this table, is computed as the number of amino acids in the primary sequence. The remaining 25 features are as defined by Chi et al. [12]. These incorporate features of geometrical model as well as stereochemical nature of a protein's tertiary structure.

Table 1 Serial numbers, as successive components of the feature vector $\underline{X}$, of features are shown as "1)", "2)", etc. in this table. Local feature numbers 2 to 17 are *histogram features* and global features 18 to 26 are *texture measures* of pixel matrix. The abbreviations in parenthesis for each feature are used throughout the text

Features	
Local	Global
16 Histogram feature	
Band1	1) Length (Len)
2) Histogram [1, 1] (H1)	
3) Histogram [1, 2] (H2)	9 Texture measure
4) Histogram [1, 3] (H3)	Orderliness group
5) Histogram [1, 4] (H4)	18) Maximum probability (Mxpr)
Band2	19) Uniformity Of energy (Ener)
6) Histogram [2, 1] (H5)	20) Entropy (Entr)
7) Histogram [2, 2] (H6)	Contrast group
8) Histogram [2, 3] (H7)	21) Homogeneity (Homo)
9) Histogram [2, 4] (H8)	22) Contrast (Cont)
Band3	23) Dissimilarity (Dis)
10)Histogram [3, 1] (H9)	24) Inverse difference moment (Idm)
11)Histogram [3, 2] (H10)	Statistical group
12)Histogram [3, 3] (H11)	25) Cluster tendency (Clust)
13)Histogram [3, 4] (H12)	26) Correlation (Cor)
Band4	
14)Histogram [4, 1] (H13)	
15)Histogram [4, 2] (H14)	
16)Histogram [4, 3] (H15)	
17)Histogram [4, 4] (H16)	

As no web-server or software is available for computing these features, we have developed our own programs on Linux platform to compute these features, as described in Chi et al. [12]. The names and notations of these features are retained as in their paper. Among these, there are a total of 16 local features (*histogram features*) and the remaining nine are global features that are measured as *texture measures*. All these computed from the pixel matrix of inter-residue distances. If only specific (structural) domain of a protein is under consideration then the feature vector is computed only for that portion.

Pixel Matrix Pair-wise Euclidean distances between the coordinates of the backbone residues of the protein under consideration are computed. (This matrix is symmetric with diagonal elements as zeros. So, only its upper or lower triangle is computed). This inter-residue distance matrix is converted into a Pixel Matrix where distances are converted to 32 *gray levels*: minimum distance = 0 and maximum distance = 31 pixels.

The 16 local (histogram) features are obtained as follows. The pixel matrix is partitioned diagonally into four band-strips as illustrated in Fig. 1. In each band, four local features are computed as relative frequencies of inter-residue distances in the (pixel) ranges 0 to 7; 8 to 15; 16 to 23; and 14 to 31.

The nine global features are calculated as *texture measures* of the pixel matrix; these are defined as functions

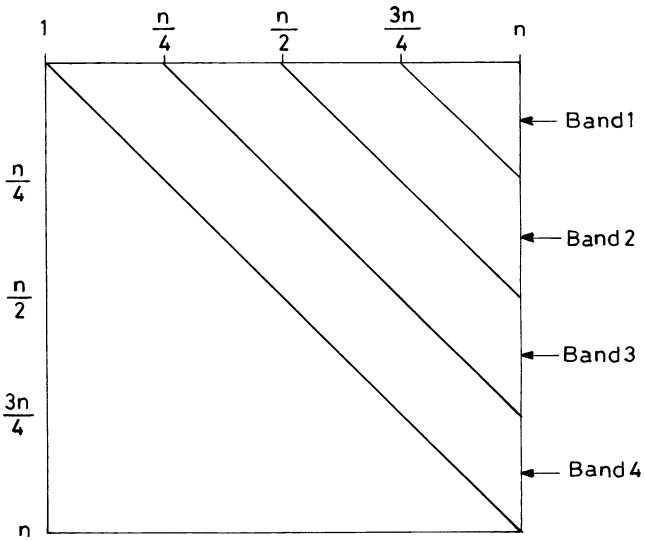

Fig. 1 Illustration of four bands in an *n*X*n* pixel matrix; values above the top horizontal boundary indicate column nos; and those on the left of 1st vertical boundary denote the row nos. Pixel at *i*th row, *j*th column corresponds to distance between *i*th & *j*th residues

Table 2 Cumulative percentage of variance contributed by the first five PCs in different classes

	All Alpha	All Beta	Alpha/Beta	Alpha+Beta
PC 1	49.21	45.60	48.90	41.80
PC 2	15.49	17.41	15.31	25.07
PC 3	11.11	10.67	10.57	10.55
PC 4	6.23	7.41	8.10	6.11
PC 5	3.54	4.88	5.17	2.69
Total	85.58	85.97	88.05	86.22

of the spatial variation in pixel intensities (*gray levels*). These are computed using the *gray level co-occurrence matrix* (GLCM), which explains the distribution of a pairs of gray levels in the pixel matrix. The (i, j)th element of the GLCM denoted by $P(d, \theta)$ is computed as the number of times the gray level i and j are separated by distance 'd' with direction 'θ' in the pixel matrix. In our computations, we have taken $d=1$ and $\theta=(0, 45, 90, 135, 180, 225, 270, 315)$. We thus obtain eight GLCM matrices in total.

The desired nine *texture measures* are computed using the formula given in Chi et al. [12]. Our computer program to calculate the feature vector may be obtained from the corresponding author.

Pixel matrix and local structural folds Pair-wise distances between C_α backbone residues are of key importance in determination or prediction of protein structures — especially the secondary structure and local folds of the tertiary structures [15]. The ab initio methods of prediction of protein tertiary structure from primary sequence extensively rely upon inter-residue distances. Conventional statistical estimates of the lower and upper bounds on inter-residue distances in *alpha-helix*, *beta sheets*, and *coils* obtained from large samples, are often useful for short range span: For example, if amino acid a primary sequence positions 'i' and 'j' are both part of an alpha helices fold in the tertiary structure then the distance d_{ij} (i.e., distance between them in 3-dimensional Euclidean space) between them would satisfy, $d_{ij} \in [4.5, 7.5]$ if j is 3rd or 4th neighbor of 'i' on the primary sequence, etc. However, no such estimates are available for medium or long-range spans in general, e.g., for $j>i+20$, etc. Different methods deploy different approaches to compute/estimate or otherwise incorporate inter-residue distances; for example, lattice models [16],

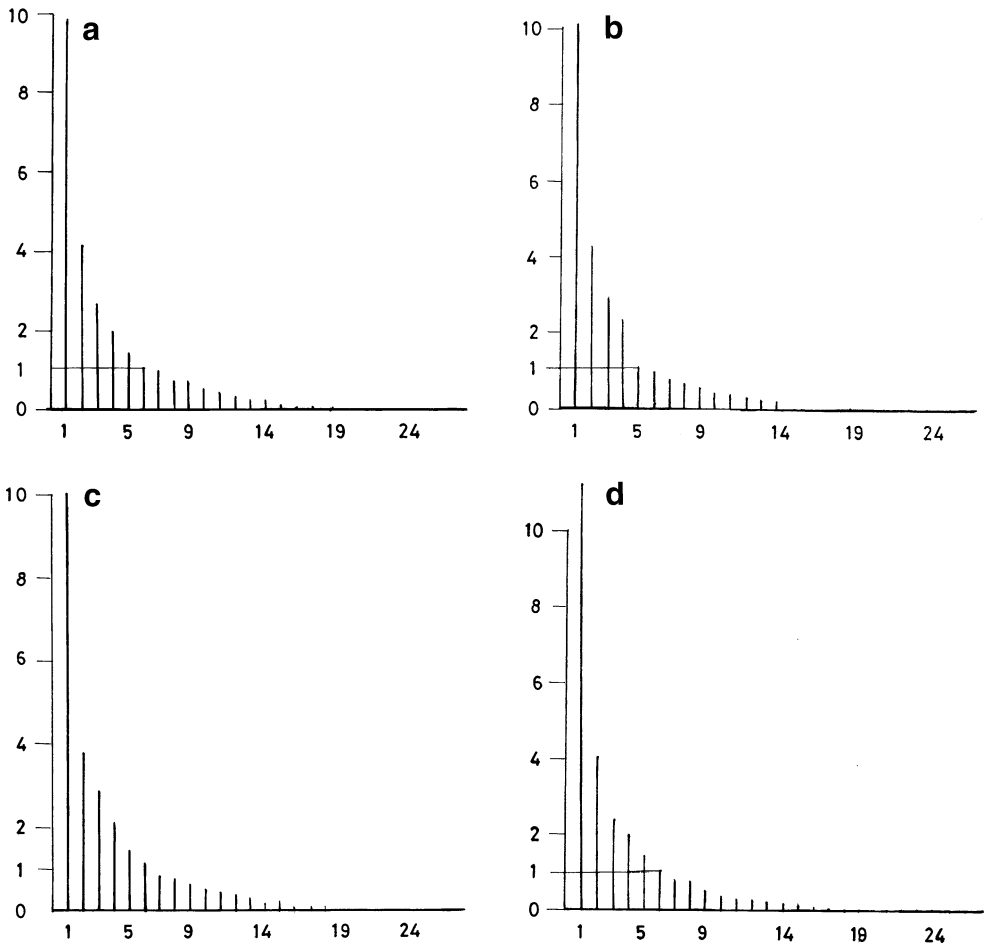

Fig. 2 The bar-diagrams correspond to the data from class (**a**) *All Alpha*; (**b**) *All Beta*; (**c**) Alpha/Beta; and (**d**) *Alpha +Beta*. In each diagram, labels, 1, 2, …, etc on the *X*-axis denote the successive *principal components* PC1, PC2, ….,etc. The *Y*-axis shows *eigenvalues* of covariance matrix of the 26 dimensional feature vector. A horizontal line is drawn at *eigenvalue*=1 for clear indication of the fact that in each class, the *eigenvalue* corresponding to the first five PCs is >1. In most cases the *eigenvalues* corresponding to PC16 onward are negligible

Table 3 The features that were found important in terms of statistically significant (confidence level >90%) correlation with the first three* PCs are listed here for the data described in section "Data set for common structural fold within a class"; abbreviated names of features are as in Table 1. (*correlation with other PCs are not found significant). Magnitude of correlation coefficient in each case is ≥0.75. Superscript '(−)' indicates that its sign is negative

Class	Significant features
All Alpha	H2, $H5^{(-)}$, $H9^{(-)}$, $H10^{(-)}$, Ener, $Entr^{(-)}$, Homo, $Cont^{(-)}$, $Dis^{(-)}$, Idm, Cor
All Beta	$Len^{(-)}$, $H1^{(-)}$, $H5^{(-)}$, H9, $Ener^{(-)}$, Entr, $Homo^{(-)}$, Cont, Dis, $Idm^{(-)}$, $Cor^{(-)}$
Alpha/Beta	H1, $H8^{(-)}$, $H9^{(-)}$, $H10^{(-)}$, $H11^{(-)}$, H12, H13, Ener, $Entr^{(-)}$, Homo, $Cont^{(-)}$, $Dis^{(-)}$, Idm, Cor, Mxpr
Alpha+Beta	$Len^{(-)}$, $H1^{(-)}$, $H2^{(-)}$, $H8^{(-)}$, H9, H10, H12, H16, Mxpr, $Ener^{(-)}$, Entr, $Homo^{(-)}$, Cont, Dis, $Idm^{(-)}$, Clust, $Cor^{(-)}$

threading [17], and/or nonparametric statistics and knowledge-based heuristic [10].

The bands in pixel matrix incorporate important information on inter-residue distance distribution in certain structural folds. In view of the earlier studies [18], if the pixel matrices of *alpha helices* in proteins of length n are aligned then there will be maximal alignment and matching in the segments (in one or more of the four bands) that are close and parallel to the diagonal. Thus, for *helices* of length $\leq n/4$, the value of feature H1 will be almost the same in all the corresponding feature-vectors and H2 may also have small variance in any sample of these feature-vectors.

For *parallel beta sheets* the aligned portions of pixel matrices would be away from the diagonal in the bands corresponding to the size of the sheet. Thus, for example, features like H4 and H8 and may be H3, H7 would have small variances in the sample of feature-vectors of parallel *beta sheets* of length greater than $n/4$ and $\leq n/2$. Alignment of inter-residue distances for *anti-parallel beta sheets* would span across segments perpendicular to the diagonal of their superimposed pixel matrices. These segments would be spread across one or more bands depending upon the length of the anti-parallel beta sheets. Thus, the distribution of pixels and the angle between the farther ones in these segments would be similar across the motifs (aligned portions) of such sheets.

In essence, the length of protein, 16 *histogram*-features, and *texture measures* depending upon 'd ' and the direction *angle* 'θ ' of corresponding GLMCs P(d, θ), would collectively extract the secondary structural (local) folds of different types and sizes and their relative and interactive positions in the tertiary structural domains.

Structural classes and fold types

We focus on classification of protein tertiary structures in four major families — *All Alpha*, *All Beta*, *Alpha/Beta* and *Alpha+Beta*. Introduction to these structural folds with illustrative graphics may be found in [15] and in structural domain definitions of SCOP database (http://scop.mrc-lmb.cam.ac.uk/scop/). We have carried out quantitative representation and analysis in both the cases — (i) classification among these four classes (families) while considering protein domains having common fold types within a class; (ii) classification while allowing different structural folds within each class.

Data set for common structural fold within a class

Considering that SCOP database does finer structural classifications at different fold levels and is also the basis/yardstick of test of the work reported by Chi et al. [13], we have considered structural families and fold types of protein (domains) as identified in this database. For exhaustive search we randomly selected maximum possible number of high-resolution structures of proteins the structural domains of which are authenticated in SCOP such that a comparable number of non-redundant observations are available from each of the four classes of interest and such that samples from each class will contain different possible sizes and orientation of the structural domain it represents.

Development of any data-mining algorithm for predictive applications requires the data set to be bias-free. Considering this, from among the randomly selected set we have chosen a sample of 225 proteins, which are mutually non-homologous [1]. List of these with indication of specific chains and structural domains as tagged in SCOP is given in the Appendix. Pair-wise sequential homology between these was tested using ClustalW program [19] and is found to be less than ≤ 25% with most pairs having less than 18% identity.

Table 4 Coefficients (i.e., components of vectors β_j in model-Eq. 1 for jth class) of the PCs, and intercept (α_j), in logistic regression model

Regressor variable	Class		
	All Alpha	All Beta	Alpha/Beta
PC1	−0.3822	0.2904	1.5661
PC2	−1.1679	0.6538	−6.2338
PC3	−0.6	1.7431	1.8762
PC4	2.3261	−1.3407	2.4824
PC5	2.291	−0.0603	4.3087
Intercept	1.4423	−2.3951	−13.519

Table 5 Average accuracy parameters (in %): True positives (TP), false positives (FP) and area under the RO- curve (A_{ROC})

Class	TP	FP	A_{ROC}
All Alpha	75.7	16.1	88.5
All Beta	69.6	19.4	71.6
Alpha/Beta	79.7	8.3	89.7
Alpha+Beta	70.4	16.2	76.5

Common fold types within the classes of interest are: fold "a.4" of class *All Alpha* ; fold "b.1" of class *All Beta* ; fold "c.1" of class *Alpha/Beta* and fold "d.58" of class *Alpha+Beta*.

A *Jackknife* type technique is applied for optimal *training* and *cross-validation* [20, 21]. In each experiment, a random subset of the above described set of 225 proteins is used as the *training* sample and the remaining as *validation*. Everytime, the *training* sample has about 40 representatives from each class.

Data set for different structural fold types within a class

We have extended the above work on different folds within each class. This data set consists of vectors of about 30–35 proteins from each major fold type in each class. A list of these is also given in the Appendix. Structural domains satisfying non-homology at sequential levels and different structural fold types (as identified in SCOP database) are considered. The following are the different fold types chosen from the four classes of interest.

Class	Fold types considered in our study
All Alpha	Alpha Alpha Superhelix (a.207); EF hand like (a.51); DNA/RNA 3 helical (a.8); Cytochrome c (a.7);
All Beta	Concanavaline (b.51); Immunoglobin like (b.1); OB folds (b.71); Trypsin like serine protease (b.80)
Alpha/Beta	Flavodoxin (c.27); Ribonuclease H like (c.77); Thioredoxin (c.68); Tim beta (c.1)
Alpha + Beta	Beta grasp (d.30); Cystatin like (d.34); Protein kinase (d.300); Ferredoxin (d.129)

We consider classification into different fold types within each structural class. This is further extended on a combined sample for classification among the four classes, using an equal number of observations on each type of fold from a class as representative of that class.

Quantitative representation and dimensionality reduction

The 26 features listed in Table 1 are computed for the chosen dataset using our programs [22, 23] on Linux platform with the support of *bio3d* utility of R-software. *Principal component analysis* (PCA) is then applied to reduce dimension of the 26-dimensional feature vector.

Principal component analysis (PCA)

Dimensionality reduction is most sought of in mining, analysis and applications of multidimensional data. PCA is a theoretically sound method that offers dimensionality reduction while also preserving all the significant information contained in the original data. It is a method of dimensionality reduction in multivariate statistics that transforms a number of possibly correlated variables into a smaller number of mutually uncorrelated variables called *principal components*. The *k principal components* of a *k*-dimensional feature vector $\underline{X}$ are obtained by orthogonal linear transformation: *i*th *principal component* of $\underline{X}=(\underline{\nu}_i)^T \underline{X}$; where superscript 'T' denotes transpose of a vector; $\underline{\nu}_I$ denotes the *eigenvector* corresponding to the *i*th (in descending order of magnitude) *eigenvalue* of the covariance matrix of $\underline{X}$.

Multivariate statistics theory [24] shows that the first *principal component* captures maximum variability in the data, followed by the second principal component and so on. So, the first few *principal components* would provide most of the useful information contained in any random sample of observations on $\underline{X}$. Thus, for further application, instead of using *k*-dimensional vector $\underline{X}$ we may use a k^*-dimensional vector ($k^*<k$) of the first k^* *principal components* of $\underline{X}$.

As presented in section "Results" below, in our study use of only the first five (i.e., $k^*=5$) *principal components* of

Table 6 The features that were found important in terms of statistically significant (confidence level>90%) correlation with the first three* PCs are listed here for the data described in Sect. "Data set for different structural fold types within a class"; abbreviated names of features are as in Table 1. (*correlation with other PCs are not found significant). Magnitude of correlation coefficient in each case is ≥0.75. Superscript '(−)' indicates that its sign is negative

Class	Significant features
All Alpha	H1, H2, $H5^{(-)}$, $H6^{(-)}$, $H9^{(-)}$, $H10^{(-)}$, H11, Mxpr, Ener, $Entr^{(-)}$, Homo, $Cont^{(-)}$, $Dis^{(-)}$, Idm, Clust, Cor
All Beta	H1, H2, $H5^{(-)}$, $H11^{(-)}$, $H12^{(-)}$, Mxpr, Ener, $Entr^{(-)}$, Homo, $Cont^{(-)}$, $Dis^{(-)}$, Idm, Clust, Cor
Alpha/Beta	Len, $H1^{(-)}$, $H2^{(-)}$, $H3^{(-)}$, H5, H9, H10, H11, H14, H16, $Mxpr^{(-)}$, $Ener^{(-)}$, Entr, $Homo^{(-)}$, Cont, Dis, $Idm^{(-)}$
Alpha+Beta	Len, $H1^{(-)}$, $H3^{(-)}$, $H8^{(-)}$, $H9^{(-)}$, H11, H12, H13, H14, $Mxpr^{(-)}$, $Ener^{(-)}$, Entr, $Homo^{(-)}$, Cont, Dis, $Idm^{(-)}$, Clust, $Cor^{(-)}$

Table 7 Estimates of intercept and coefficients (β_j for jth PC) for different fold types in the structural class *All Alpha*. Fold type *Cytochrome-C* is used as a reference in logistic regression model

Fold type	Intercept	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Alpha Alpha superhelix	15.172	−7.716	4.1634	1.1664	−3.479	3.5792	2.882	−4.077
EF hand like	18.233	−6.812	3.3299	2.0851	−2.852	2.048	2.398	−4.273
DNA/RNA 3 helical	18.448	−6.605	2.8673	2.5311	−2.704	1.4128	2.620	−4.111

the 26-dimenstional feature vector of protein-structure is found adequate.

Relation of PCs with original descriptors

There need not be a one-to-one correspondence between an original feature and a *principal component*. By definition, every *principal component* being a linear combination of original features would represent their combined effect. First few *principal components*, which explain maximum variability (and hence the information content) of the data would capture the joint effect of the important features and thus preserve the collective role of original descriptors more efficiently.

In section "Materials and methods" we have highlighted the importance of pixel matrix and hence the feature vectors vis-à-vis the protein's secondary structural folds. While some individual *histogram features* might capture the signature (motif) of an *alpha helix* or *beta sheet* of specific lengths, the *anti-parallel beta sheets* require several global features as well. As a single protein could have several *local folds* of varied sizes at different positions, collective role of all the features is essential to represent these. Even if single structural domains per protein are considered, there would be diversity of sizes and relative positioning across the *training* sample from which the characteristic of a class is to be extracted.

Therefore, the projection of original data into a reduced dimensional space is required to be such that the collective role of all the features is reflected. *Principal component analysis* fulfils this requirement with an additional advantage that the sign and magnitude of the correlation coefficients of different features with a *principal component* also reflect their relative importance in representing the data.

Classification using multi-class logistic regression

Consider the problem of classifying a feature vector $\underline{Y}$ in one of the C classes of interest. A standard multi-class logistic regression model defines the probability p_j of $\underline{Y}$ belonging to jth class, j=1, 2, …., C-1 as a *logit* function [25]:

$$\ln\left(\frac{p_j}{1-p_j}\right) = \alpha_j + \underline{Y}^T\underline{\beta}_j + \text{random error term.} \quad (1)$$

The probability of $\underline{Y}$ belonging to the Cth class is defined as $p_j = 1 - \sum_{j=1}^{C-1} p_j$. This class is termed the reference class.

Fitting of such a model amounts to estimating the intercepts α_j and the vector $\underline{\beta}_j$ of unknown coefficients using a *training* sample — of observations (on $\underline{Y}$) from the C classes of interest, so as to minimize the squared sum of random error. Once the model is fitted, any given vector $\underline{Y}$ is assigned to the class to which it would lie with maximum probability.

In our study, C=4. Having estimated the *principal components* of the 26-dimensional feature vector $\underline{X}$, we obtain for each observation ($\underline{X}_i$; i=1, 2, …, n) in the *training* sample, the corresponding vector ($\underline{Y}_i$) of the first five *principal components* and fit the *logistic regression* model.

Results

We first present the results for the data set described in section "Data set for common structural fold within a class".

Principal component analysis

We found that in each of the four classes of interest, the first five *principal* components explain nearly 85% (see

Table 8 Estimates of intercept and coefficients (β_j for jth PC) for different fold types in the structural class *All Beta*. Fold type *Trypsin-like-serine-protease* is used as a reference in logistic regression model

Fold type	Intercept	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Concanavaline	111.381	38.774	22.323	−50.104	39.854	99.182	1.2854	−5.268
Immunoglobin like	172.176	57.598	39.323	−73.44	81.096	28.744	47.703	6.078
OB folds	171.860	57.325	39.482	−73.679	81.229	27.734	47.809	5.165

Table 9 Estimates of intercept and coefficients (β_j for *j*th PC) for different fold types in the structural class *Alpha/Beta*. Fold type *Tim-Beta* is used as a reference in logistic regression model

Fold type	Intercept	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Flavodoxin	146.439	53.569	−46.669	35.449	−143.669	−24.355	48.288	5.488
Ribonuclease H like	147.600	53.694	−46.664	36.826	−141.781	−24.945	48.642	5.819
Thioredoxin	147.571	53.645	−47.174	36.862	−142.458	−24.765	48.252	4.543

Table 2) of the total variance in the *training sample*. Figure 2 shows the significance of the first five *principal components* (PCs).

Role of different features in structural motifs

Analysis of the *correlation* of the first five PCs with different features shows significant difference in influence of certain features in different structural classes – in terms of statistical significance of the correlation coefficient and its magnitude and/or direction (positive or negative). Table 3 summarizes the main results.

From, this table it is clear that the classes *All Beta* and *Alpha+Beta* are more similar to each other as compared to the other two classes and the classes *All Alpha* and *Alpha/Beta* are similar with respect to the features that are found to describe them. *Len* (length of the protein sequence under consideration) is found as a significant feature in the description of the *All Beta* and *Alpha+Beta* but not in the other two classes; another global feature *mxpr* (maximum probability) is found as significant only in *Alpha/Beta* and *Alpha+Beta*. Histogram feature H2 is found significant only in *All Alpha*; H13 in only *Alpha/Beta*; and H16 only in *Alpha+Beta*.

Cluster tendency (*Clust*) is found significant only in describing the class *Alpha+Beta*. Most of the other *texture measures* are found significant in all the classes except that the signs of their correlation with the combined descriptors (the first three PCs) are opposite in the *All Alpha* and *Alpha/Beta* classes as against *All Beta* and *Alpha+Beta*.

Predictive classification

As described earlier, several computational experiments are conducted using random subsets of the dataset described in section "Data set for common structural fold within a class" as *training* samples. In each experiment, 4-class logistic regression is fitted using the R-software (http://www.r-project.org/); the first five *principal components* (PC) are regarded as the explanatory (*regressor*) variables. *Alpha+Beta* class is considered as the reference class. Classes of the feature-vectors in the *validation* samples are predicted using the fitted model.

Coefficients of the PCs in this model are shown in Table 4. The best model gave more than 82% prediction accuracy for each class. Averages (of cross-validation results) of the accuracy parameters are shown in Table 5.

The accuracies of predictive classification by other models have also been satisfactory. The following table shows average performance.

Results for different folds within a class

For the data set described in section "Data set for different structural fold types within a class" we have found that the first five PCs explain more than 85% of variation in the data. The contributions of individual PCs are also comparable with those shown in Table 2 and Fig. 2.

Role of different features

Analysis of correlation of the first five PCs with the features described in Table 1 shows interesting results. As far as comparison between classes is concerned the roles of features significant in distinguishing between the classes remain similar to those summarized in Table 3. However, comparisons within a class show distinct roles of certain features with respect to different folds. Table 6 underneath summarizes the key results.

Table 10 Estimates of intercept and coefficients (β_j for *j*th PC) for different fold types in structural class *Alpha + Beta*. Fold type *Ferredoxin* is used as a reference in logistic regression model

Fold type	Intercept	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Beta grasp	0.293	0.201	0.336	0.528	−0.510	−1.057	−0.861	1.182
Cystatin like	1.844	0.373	1.683	0.856	2.258	1.122	0.398	1.415
Protein kinase like	−56.551	3.281	25.134	6.901	22.087	23.250	10.222	24.752

Table 11 Average accuracy parameters (in %) for classification of different fold types within class *All Alpha*: True positives (TP), False positives (FP) and area under the ROC- curve (A_{ROC})

Fold type	TP	FP	A_{ROC}
Alpha Alpha superhelix	83.9	7.5	93.2
EF hand like	57.1	14.6	74.4
DNA/RNA 3 helical	62.3	13.9	79.4
Cytochrome C	91.7	4.1	91.6

Because of higher within-class variability (due to different fold types), except H16, roles of no other *local* or *global features* are so distinct as found in the case of common fold types within a class (section "Role of different features in structural motifs" above). Except "*Len*", no other feature is found to prominently distinguish even between groups of classes. The role of length of protein sequence (*len*) is now found significant in distinguishing between the classes *Alpha/Beta* and *Alpha+Beta* against *All Alpha* and *All Beta*. This indicates that the sizes of *local* (secondary) *structural* domains are more variable with respect to the fold types in the latter classes as compared to those in the former. This is justified in view of the fact that the classes *Alpha/Beta* and *Alpha+Beta* already have a mixed kind of *local structural* domains, so variability with respect to different fold types within such a class does not influence the role of length (size) of the domains.

Predictive classification of structural fold types within a class

As described in section "Data set for different structural fold types within a class", within each class we have considered proteins with four different types of structural folds. We have used multi-class logistic regression on the first seven PCs, to predict these structural folds within each class. As in each class, the first seven PCs explained more than 85% of the total variation in data, so the first seven PCs were considered as predictor variables. Similar to the case of data with common structural folds within a class, we have carried out several computational experiments using the jackknife technique of cross-validation.

Table 12 Average accuracy parameters (in %) for classification of different fold types within class *All Beta*: True positives (TP), False positives (FP) and area under the ROC- curve (A_{ROC})

Fold type	TP	FP	A_{ROC}
Concanavaline	89.6	3.4	98.6
Immunoglobin like	66.7	12.2	90.1
OB folds	59.4	13.1	85.9
Trp like serine protease	88.9	3.8	98.2

Table 13 Average accuracy parameters (in %) for classification of different fold types within class *Alpha/Beta*: True positives (TP), false positives (FP) and area under the ROC- curve (A_{ROC})

Fold type	TP	FP	A_{ROC}
Flavodoxin	85.3	12.1	90.2
Ribonuclease H like	59.6	12.4	84.6
Thioredoxin	61.4	10.6	82.3
Tim Beta	89.3	1.6	97.4

The estimated regression coefficients and intercepts of best models for each class under consideration are shown in Tables 7, 8, 9, and 10. For each class, the models show overall predictive accuracy (i.e., percentage of correctly classified fold types) ≥ 73%. Averages (of cross-validation results) of the accuracy parameters are shown in Tables 11, 12, 13, and 14.

Predictive classification of using different structural folds within a class

We have also carried out computational experiments on predictive classification by multi-class logistics using *training* samples of sizes about 40 from each of the structural classes – *All Alpha*, *All Beta*, *Alpha/Beta*, and *Alpha+Beta*. In this case the first seven PCs explain the desired (> 85%) of total variation in the data. In all experiments, the *training* sample from a class consists of about ten observations for each of the four different types (described in section "Data set for different structural fold types within a class") of folds prominently found in this class. Class *Alpha+Beta* is regarded as the reference class for fitting of the logistic regression model with the first seven PCs as the predictor variables.

Estimated parameters of the model are shown in Table 15 and average (of cross-validation results) accuracy results are shown in Table 16.

The overall accuracy of correct classification (TP) in the best model is around 74%. This as well as the average TP for each class are lower as compared to those for the case (section "Predictive classification" above) when the *training* sample from a class consisted of common structural

Table 14 Average accuracy parameters (in %) for classification of different fold types within class *Alpha+Beta*: True positives (TP), false positives (FP) and area under ROC- curve (A_{ROC})

Fold type	TP	FP	A_{ROC}
Beta grasp	64.7	18.1	86.5
Cystatin like	74.2	8.2	92.4
Protein kinase like	98.4	1.1	99.2
Ferredoxin	62.3	9.4	86.9

Table 15 Coefficients (components of vectors β_j in model-equation (1) for *j*th class) of the PCs; and the intercept (α_j)

Class	Intercept	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Alpha	−0.661	0.065	0.897	0.542	−1.498	−0.872	1.418	0.359
Beta	−0.18	0.164	−0.379	0.476	0.111	−0.711	−0.385	−2.099
Alpha/Beta	0.097	0.232	0.059	0.082	−1.184	−1.637	0.141	−2.038

fold type. It is expected because in the present case the size of the *training* sample is comparable to that used in the case of common with-in class folds, but this *training* sample is significantly heterogeneous.

Discussion

Statistical modeling and analysis of protein data carried out in this paper has provided important quantitative insight into major structural families (as identified in SCOP database) and has also offered computationally feasible and efficient predictive methods for their classification. Computational methods using feature vectors are remarkably simpler to structural homology for classification of proteins. Our approach has added advantages of reduced dimension of the feature vector and use of statistical data mining.

It is notable that though we have reduced the dimension of quantitative feature-vector representation of protein tertiary structures to at the most seven, the accuracy of structural classification we get is comparable to or better than that of Chi et al. [12, 13]. In the case of common fold types representing a structural class, the dimension as less than five is adequate for predictive classification with high accuracy. Apart from dimensionality reduction, insight into relative importance of certain features in specific structural folds is another gain over the best-known relevant approach [12].

Efficient and theoretically sound method of *principal component analysis* (PCA) is used here for dimensionality reduction. *Principal components* being linear transformations of the original data are easy to compute. Moreover, these being orthogonal (and hence uncorrelated) to each other can also be used as explanatory variables in the powerful predictive applications of regression modeling.

Table 16 Average accuracy parameters (in %): True positives (TP), false positives (FP) & area under ROC- curve (A_{ROC})

Class	TP	FP	A_{ROC}
All Alpha	59.6	12.5	83.7
All Beta	67.8	16.8	82.6
Alpha/Beta	57.7	18.5	77.4
Alpha+Beta	69.5	5.2	88.7

Comparative analysis in terms of significant correlation of features with the key *PC*s reveals interesting results on relative importance and representative roles of certain topological, structural and stereochemical features in describing and distinguishing the four major 'classes' of protein structures.

As shown in Table 3, no histogram features in band4, i.e., no long-range inter-residue distances are important in characterizing the *All Alpha* and *All Beta* type folds. *Texture measures* and hence topological as well as stereochemical factors are found more important (though mostly with respect to the sign of correlation with the important PCs) than *local features* in distinguishing between these classes. *Alpha/Beta* structures appear closer to *All Alpha* with respect to these features, whereas *Alpha+Beta* types seem to share this similarity with *All Beta*.

When common fold types within a class are considered, *length* (*len*) of a protein sequence under consideration is found to play an important role in distinguishing *All Beta* and *Alpha+Beta* classes from *All Alpha* and *Alpha/Beta*. Another global feature *mxpr* (maximum probability) is found to distinguish *Alpha/Beta* and *Alpha + Beta* from the other two classes. Interestingly, for each class, one significant *local feature* or *global feature* along with the above is also found as an important descriptor. It is notable that within class variability different fold types perturb this influence except for the roles of H16 and *len* in the mixed class *Alpha + Beta*.

Exact values of correlation coefficients and the regression coefficients of the PCs in each class can be used for detailed statistical analysis of interactive roles of local folds in a tertiary structure, which is not possible otherwise. Using these values, computer aided molecular designs of certain structures – e.g., functionally important tertiary motifs – may be obtained. Random variation in values of features found important in distinguishing different types of structural folds (e.g., Table 6) would provide computationally simpler techniques than molecular dynamics for simulation of protein tertiary folds and would also help in testing the empirical hypotheses on this yet un-deciphered phenomenon. We shall report some results in this regard subsequently.

Multi-class logistic regression has been extensively used in wide-ranging applications including medical- and bio-informatics and immunology (e.g., [25, 26]). Here it provides a computationally feasible and predictive method

of classifying protein structural families. It is remarkably simpler in computation than the methods of structural homology used to distinguish between structurally similar and dissimilar proteins. Another significant importance of this method lies in the fact that we can assign confidence levels of accuracy to predictive classifications and also to class-definition in terms of the feature vectors.

Our results for classification between four major structural classes, with (i) common fold types representing a class; and also for (ii) different fold-types within a class, are excellent in terms of overall accuracy of classification and area under ROC. Often in predictive applications, there is compromise between true positives and false positives. A_{ROC} — area under the receiver operating characteristic curve (ROC) provides a comprehensive measure of reliability and consistency of a predictive method or model [27, 28]. The models fitted here for classification into one of the four structural classes and those for further discrimination among different fold types within a class are found to be good in terms of this criterion. The corresponding regression models can be used in predictive application to classify any new protein.

Further, the present study strengthens the possibility of deploying similar quantitative modeling to predict functionally important structural motifs or functional sites in proteins. We have used it to infer the presence and location of certain functional sites in new or predicted structures of proteins [29].

Acknowledgments This work was carried out at the Bioinformatics Centre, University of Pune, and at the Indian Institute of Technology Bombay. The authors are thankful to these institutions for the infrastructural and administrative support.

Appendix

List of proteins referred to in section "Materials and methods"

(a) pdb ids of the 225 proteins in the data set referred in section "Data set for common structural fold within a class"

All Alpha	All Beta	*Alpha/Beta*	*Alpha+Beta*
1aoy	1a3r	1a4m	1ab8
1b9m	1b4r	1aj2	1afj
1bby	1bww	1b5t	1aop
1bia	1cf1	1bd0	1b64
1bja	1cvr	1bqg	1cg2
1bl0	1dqi	1ccw	1dur
1bm9	1ehx	1ct5	1ekr
1cf7	1ex0	1d8c	1f0x
1d8j	1ex0	1d8w	1f3v
1dp7	1f00	1dbt	1f9y
1e17	1gof	1dos	1feh
1e3h	1gyv	1dxe	1ffg
1ef4	1i8a	1e4m	1fi4
1etx	1ifr	1eep	1ftr
1fc3	1im3	1egv	1fvg
1fli	1jz8	1ejx	1gmu
1fp1	1kmt	1ezw	1gpj
1fse	1kyf	1f6y	1h72
1fsh	1l6p	1fcq	1hbn
1g3w	1lla	1frb	1hbn
1gvd	1lmi	1gkp	1hw8
1hc8	1m1x	1h41	1i19
1hcr	1msp	1hl9	1i1g
1hks	1n9p	1i0d	1in0
1hlv	1nci	1i60	1iuj
1hst	1nep	1itu	1ivz
1hw1	1o6v	1j5s	1j27
1i1g	1o75	1j6o	1j5e
1i27	1osy	1j79	1jmt
1i5z	1p7h	1jfx	1k47
1if1	1pby	1jqn	1kkh
1ign	1pl3	1jub	1kn6
1irz	1q0e	1k77	1koh
1ixc	1qfh	1kbl	1kp6
1ixs	1r4x	1lt8	1l3k
1j5e	1roc	1luc	1lou
1jgs	1svb	1m5w	1lq9
1jhf	1tza	1n8f	1lxn
1jhg	1u2c	1nfp	1m1h
1k6y	1uad	1nqk	1mg7
1k78	1ug9	1nth	1mla
1kqq	1v8h	1nvm	1mli
1ku9	1vca	1o1z	1mwq
1l8q	1vca	1ob0	1nh8
1ldd	1xak	1ohl	1nue
1lva	1xo8	1olt	1nxi
1mkm	2a9d	1onw	1nza
1mzb	2b20	1oy0	1o51
1o57	2c9q	1p1m	1o8b
1ofc	2dpk	1p1x	1oy8
1okr	2h7w		1pbu
1opc	2hft		1pca
1oyw	2j2z		1phz
1p7i	2mcm		1pie
1pp7	4kbp		1pys
1q1h			1q4r
1r1t			1q5y
1r71			1q8b
1r7j			1q8k
1rep			1qd1

(b) SCOP ids of the 225 proteins listed in the above Table

All Alpha	*All Beta*	*Alpha/Beta*	*Alpha+Beta*
d1aoya__16087.pdb	d1a3rl2_20890.pdb	d1a4ma__29014.pdb	d1ab8a__39414.pdb
d1b9ma1_16118.pdb	d1b4ra__22072.pdb	d1aj2a__29665.pdb	d1afja__39338.pdb
d1bbya__16149.pdb	d1bwwa__20518.pdb	d1b5ta__29676.pdb	d1aopa1_39501.pdb
d1biaa1_16083.pdb	d1cf1a1_21907.pdb	d1bd0a2_28642.pdb	d1b64a__39306.pdb
d1bjaa__16122.pdb	d1cvra1_21949.pdb	d1bqga1_29217.pdb	d1cg2a2_39360.pdb
d1bl0a1_16053.pdb	d1dqia__22357.pdb	d1ccwb__29646.pdb	d1dura__38943.pdb
d1bm9a__16116.pdb	d1ehxa__21950.pdb	d1ct5a__28663.pdb	d1ekra__39380.pdb
d1cf7a__16151.pdb	d1ex0a1_90465.pdb	d1d8ca__29325.pdb	d1f0xa1_39483.pdb
d1d8ja__16153.pdb	d1ex0a2_90466.pdb	d1d8wa__29394.pdb	d1f3va__39382.pdb
d1dp7p__16159.pdb	d1f00i1_22368.pdb	d1dbta__28539.pdb	d1f9ya__83249.pdb
d1e17a__16143.pdb	d1gofa1_21807.pdb	d1dosa__29175.pdb	d1feha3_38998.pdb
d1e3ha1_16257.pdb	d1gyva__70790.pdb	d1dxea__29310.pdb	d1ffgb__39384.pdb
d1ef4a__16272.pdb	d1i8aa__61951.pdb	d1e4mm__59226.pdb	d1fi4a2_59848.pdb
d1etxa__18978.pdb	d1ifra__71203.pdb	d1eepa__28636.pdb	d1ftra1_39485.pdb
d1fc3a__16237.pdb	d1im3d__62568.pdb	d1egva__29652.pdb	d1fvga__39408.pdb
d1flia__16160.pdb	d1jz8a1_67830.pdb	d1ejxc2_83185.pdb	d1gmua2_65336.pdb
d1fp1d1_59939.pdb	d1kmta__77442.pdb	d1ezwa__29558.pdb	d1gpja3_65453.pdb
d1fsea__60000.pdb	d1kyfa1_73220.pdb	d1f6ya__29673.pdb	d1h72c2_60713.pdb
d1fsha__60006.pdb	d1l6pa__73626.pdb	d1fcqa__65006.pdb	d1hbna2_60899.pdb
d1g3wa1_65133.pdb	d1llaa3_21861.pdb	d1frba__28665.pdb	d1hbnc__60902.pdb
d1gvda__83338.pdb	d1lmia__78098.pdb	d1gkpa2_70232.pdb	d1hw8a1_61298.pdb
d1hc8a__70963.pdb	d1m1xa1_74422.pdb	d1h41a1_83472.pdb	d1i19a1_61522.pdb
d1hcra__16020.pdb	d1mspa__22333.pdb	d1hl9a2_90651.pdb	d1i1ga2_65983.pdb
d1hksa__16172.pdb	d1n9pa__80343.pdb	d1i0da__61487.pdb	d1in0a1_83694.pdb
d1hlva1_65854.pdb	d1ncia__22191.pdb	d1i60a__71118.pdb	d1iuja__90701.pdb
d1hsta__16140.pdb	d1nepa__80440.pdb	d1itua__71423.pdb	d1ivza__76863.pdb
d1hw1a1_16111.pdb	d1o6va1_81099.pdb	d1j5sa__71580.pdb	d1j27a__90778.pdb
d1i1ga1_65982.pdb	d1o75a1_81117.pdb	d1j6oa__77088.pdb	d1j5ej__71553.pdb
d1i27a__61555.pdb	d1osya__93502.pdb	d1j79a__62675.pdb	d1jmta__63180.pdb
d1i5za1_83669.pdb	d1p7hl1_94271.pdb	d1jfxa__62943.pdb	d1k47a2_72041.pdb
d1if1a__16183.pdb	d1pbya3_94419.pdb	d1jqna__77159.pdb	d1kkha2_72646.pdb
d1igna1_16048.pdb	d1pl3a__88158.pdb	d1juba__90908.pdb	d1kn6a__72770.pdb
d1irza__76772.pdb	d1q0ea__95504.pdb	d1k77a__72096.pdb	d1koha2_68720.pdb
d1ixca1_83764.pdb	d1qfha1_21893.pdb	d1kbla1_68384.pdb	d1kp6a__39397.pdb
d1ixsb1_76933.pdb	d1r4xa1_97054.pdb	d1lt8a__78186.pdb	d1ll3ka1_73539.pdb
d1j5er__71561.pdb	d1roca__97673.pdb	d1luca__29547.pdb	d1loua__39323.pdb
d1jgsa__66683.pdb	d1svba1_21814.pdb	d1m5wa__84836.pdb	d1lq9a__78129.pdb
d1jhfa1_63057.pdb	d1tzaa__107468.pdb	d1n8fa__85397.pdb	d1lxna__84737.pdb
d1jhga__19009.pdb	d1u2ca1_107610.pdb	d1nfpa__29555.pdb	d1m1ha2_78416.pdb
d1k6ya1_68239.pdb	d1uadc__88379.pdb	d1nqka__92050.pdb	d1mg7a2_84955.pdb
d1k78a1_68255.pdb	d1ug9a3_99363.pdb	d1ntha__80730.pdb	d1mlaa2_39383.pdb
d1kqqa__72885.pdb	d1v8ha1_119870.pdb	d1nvma2_86250.pdb	d1mlia__39070.pdb
d1ku9a__77544.pdb	d1vcaa1_21649.pdb	d1o1za__86555.pdb	d1mwqa__91481.pdb
d1l8qa1_77809.pdb	d1vcaa2_21685.pdb	d1ob0a2_81257.pdb	d1nh8a2_80508.pdb
d1ldda__73841.pdb	d1xaka__115037.pdb	d1ohla__87035.pdb	d1nuea__39076.pdb
d1lvaa1_74276.pdb	d1xo8a__115698.pdb	d1olta__93334.pdb	d1nxia__86381.pdb
d1mkma1_79242.pdb	d2a9da1_126431.pdb	d1onwa2_87173.pdb	d1nzaa__86444.pdb

All Alpha	*All Beta*	*Alpha/Beta*	*Alpha+Beta*
d1mzba__91497.pdb	d2b20a1_127685.pdb	d1oy0a__87543.pdb	d1o51a__92480.pdb
d1o57a1_92483.pdb	d2c9qa1_130138.pdb	d1p1ma2_87697.pdb	d1o8ba2_81181.pdb
d1ofcx2_92827.pdb	d2dpka1_131616.pdb	d1p1xa__104060.pdb	d1oy8a1_87563.pdb
d1okra__93269.pdb	d2h7wa1_136225.pdb		d1pbua__88030.pdb
d1opca__16231.pdb	d2hfta1_21951.pdb		d1pcaa1_39063.pdb
d1oywa1_93760.pdb	d2j2za1_137974.pdb		d1phza1_39358.pdb
d1p7ia__94279.pdb	d2mcma__22207.pdb		d1piea2_94707.pdb
d1pp7u__94973.pdb	d4kbpa1_22345.pdb		d1pysb4_39310.pdb
d1q1ha__95580.pdb			d1q4ra__95823.pdb
d1r1ta__104769.pdb			d1q5ya__95950.pdb
d1r71a__104823.pdb			d1q8ba__96201.pdb
d1r7ja__104836.pdb			d1q8ka2_104557.pdb
d1repc1_16125.pdb			d1qd1a1_39493.pdb

(c) pdb ids of the proteins in the data set referred to in section "Data set for different structural fold types within a class"

Codes of the fold types are as in SCOP (*c.f.* Table in section "Data set for different structural fold types within a class" for names)

All distinct structural domains (found in SCOP) of the type listed here were used in the study. Therefore no separate list with SCOP ids of those is given here. Sequential homology between most pairs is <20%. However, in some pairs within a subclass (e.g., in a.207) it is greater than 45%, common domains of both members of such pairs were not used simultaneously in *training* or *validation* samples.

All Alpha				All Beta			
a.207	a.51	a.8	a.7	b.51	b.1	b.71	b.80
1A17	1C3Y	1B1B	155C	1CPN	1CCZ	1A1D	1A7S
1BC9	1DGU	1C20	1A8C	1DYK	1CDY	1A62	1BIO
1DVP	1EH2	1E17	1C2N	1DYP	1CID	1BKB	1BRU
1EYH	1EXR	1IUF	1C52	1FNY	1E5U	1BR9	1BT7
1HH8	1FPW	1JGS	1C53	1GBG	1ESO	1D2B	1CQQ
1HXI	1GGW	1KN5	1C6R	1GNZ	1F2Q	1FL0	1DPO
1JWF	1HQV	1MGT	1C6S	1GV9	1FHG	1H9K	1EAX
1KLX	1IQ3	1MIJ	1C75	1GZC	1FNL	1I40	1ELT
1LKV	1JFK	1MZB	1CC5	1H9P	1HNF	1J6Q	1EUF
1LRV	1K95	1QNT	1CCR	1J1T	1I8A	1JB3	1EXF
1N8U	1LKJ	1R36	1CO6	1KS5	1IAM	1K0S	1GVZ
1NZN	1NCX	1RI7	1COR	1LED	1IJ9	1KHI	1H4W
1OXJ	1NX2	1RR7	1COT	1LU1	1JBJ	1KL9	1HJ8
1PAQ	1NYA	1RYU	1CTJ	1MVE	1JCV	1KRS	1HJ9
1PC2	1OOI	1S7E	1CXC	1MVQ	1JPE	1KXL	1K2I
1Q2Z	1PUL	1V3F	1CYJ	1NLS	1MFM	1Q46	1KXB
1R8M	1Q80	1WJ5	1E29	1O4Y	1OAL	1SNC	1LO6
1RW2	1QV1	1XCV	1E8E	1OA4	1OLL	1SR3	1MZA
1RZ4	1RRO	1XD7	1F1F	1OLR	1OP4	1TWL	1NN6
1TE4	1S3P	1YG2	1FI3	1S2B	1ROC	1UAP	1OP0

All Alpha				All Beta			
a.207	a.51	a.8	a.7	b.51	b.1	b.71	b.80
1WY6	1S6I	1Z91	1GDV	1SBF	1TEY	1WI5	1OS8
1XT0	1SL8	2AXL	1GKS	1UAI	1UFG	1WJJ	1P3C
1Y6I	1SNL	2BV6	1HRC	1UX6	1UGN	1X6O	1PQ7
1Y8M	1SRA	2COM	1I6D	1ZA4	1WG3	1XWE	1QNJ
1Z3X	1TUZ	2CSO	1I8O	2A6Y	1WIC	2B29	1QTF
1ZU2	1UHN	2CYY	1JDL	2A6Z	1XMW	2EIF	1RJX
2BF0	1WLM	2E34	1KX7	2AFJ	1XO8	2JA9	1SI5
2D2S	2JPO	2ESH	1LS9	2AYH	1ZXQ	2K5W	1T32
2I9C	2P71	2F5C	1MZ4	2C9A	2CU9	2PRD	1TON
2ION	2PAS	2FBH	1YCC	2ERF	2DPK	2TMP	2A31
2NSZ	2PVB	2FBI	2AI5	2NLR	2FCB	3TSS	2CXV
2O8P	2SAS	2IPQ	2C8S		2FWU		2H5C
	5PAL	2OD5	2DVH		2MFN		2RG3
		2V7F	3C2C				2SFA
		2V9V	451C				2SGA
		2VQC	5CYT				
Alpha/Beta	Alpha+Beta						
c.27	c.77	c.68	c.1	d.30	d.34	d.300	d.129
1AHN	1CXQ	1A2J	1A53	1A5R	1CEW	1A06	1FJ7
1B1A	1EH6	1BED	1CT5	1B9R	1EQK	1FOT	1J27
1CZN	1EO1	1HD2	1EDT	1E9M	1G96	1GZK	1LXJ
1DCF	1HYV	1I5G	1GQN	1EF5	1KWI	1HOW	1NO8
1DZ3	1I39	1J9B	1HW6	1GNU	1NNV	1LUF	1NZA
1EIW	1IO2	1KNG	1I60	1L2N	1Q7H	1M2R	1P1L
1F4P	1J9A	1LU4	1J5T	1MG4	1ROA	1OEC	1P1T
1FUE	1JL1	1O73	1J6O	1MJD	1SJW	1P14	1Q8B
1FYV	1MGT	1O8X	1JCM	1RAX	1SQW	1PME	1RIS
1FYX	1O13	1ON4	1K77	1RRB	1TP6	1R0P	1S79
1H05	1O1W	1PQN	1KM4	1TTN	1TUH	1RE8	1SJQ
1ID8	1OVQ	1QGV	1LYX	1UF0	1Z1S	1RJB	1UKU
1J56	1OVY	1SEN	1MXS	1V2Y	2A15	1S9J	1URR
1JBE	1P90	1UN2	1N55	1V5O	2CW9	1T46	1WEX
1M2E	1QNT	1V9W	1NFP	1WE6	2CX1	1UU3	1WEY
1MB3	1RDU	1WPI	1O1Z	1WE7	2FXT	1UV5	1WEZ
1MVO	1RIL	1XVQ	1OM0	1WF9	2GU3	1VJY	1WG5
1NAT	1SFE	1Z6M	1QWG	1WFY	2I9W	1VZO	1WI8
1NNI	1W0H	1Z6N	1SFS	1WGH	2K54	1XJD	1X4D
1P6Q	1WLJ	1ZZO	1THF	1WGK	2RFR	1XKK	1X5S
1QCZ	1YF5	2A2P	1TPE	1WGY	2STD	1XWS	1X5U
1QKK	1ZBF	2A4H	1U5H	1WI0	3B7C	1YWN	1XER
1RCF	1ZBS	2A4V	1U83	1WJ6	3CPO	1ZYL	2CPH
1RLJ	2ETJ	2B5X	1UJP	1WX7	3CU3	2B1P	2CQ0
1TMY	2GUI	2CVB	1VD6	1WXA	3DM8	2JFL	2CQ1
1YKG	2GUP	2DJK	1VPQ	1WZ0	3E99	2OJ9	2CQ3
2A9O	2HST	2DLX	1VZW	1X1M	3EBT	2PPQ	2CQB
2AYY	3E9L	2FWH	1XWY	1XO3	3EBY	2QHN	2CQG
2AYZ	3E9O	2FY6	1ZZM	1YJI	3EJV	2RG6	2CQI
2B4A		2GZP	2HVM	2BYE	3EN8	3BQC	2IVY
2C4V		2H01	2PLC	2BYF	8CHO	3C1X	2K3K

All Alpha				All Beta			
a.207	a.51	a.8	a.7	b.51	b.1	b.71	b.80
2FCR		2HFD		2C60			7FD1
2FZ5				2CR5			
2PL1				2CU1			
5NUL				2IDY			

References

1. Murzin AG, Brenner SE, Hubbard T, Chothia C. (1995) J Mol Biol 247:536–540. http://scop.mrc-lmb.cam.ac.uk/scop/
2. Orengo CA, Michie AD, Jones DT, Swindells MB, Thornton JM (1997) Structure 5:1093–1108. http://www.cathdb.info/
3. King RB (ed) (1983) Chemical applications of topology and graph theory. Elsevier, New York
4. Joshi RR, Krishnanand K (1996) J Comput Biol 3:143–162
5. Joshi RR (2001) Protein Pept Lett 8:257–264
6. Jyothi S, Mustafi SM, Chary KVR, Joshi RR (2005) J Mol Model 11:481–488
7. Thorne JL, Goldman N, Jones DT (1996) Mol Biol Evol 13:666–673
8. Chou KC (1999) J Protein Chem 18:473–480
9. Cai YD, Liu XJ, Chou KC (2003) J Comput Chem 24:727–731
10. Joshi RR, Jyothi S (2003) Comput Biol Chem 27:241–252
11. Zheng WM (2005) IJ Bioinform Res Appl 1:420–428
12. Chi PH, Scott G, Shyu CR (2005) Int J Software Engineer Knowledge Engineer 15:527–545
13. Chi PH, Shyu CR, Xu D (2006) BMC Bioinform 7:362. doi:10.1186/1471-2105-7-362
14. Jyothi S, Joshi RR (2001) Comput Chem 25:283–299
15. Branden C, Tooze J (1999) Introduction to protein structure. Garland, New York
16. Zhang Y, Kolinski A, Skolnic J (2003) Biophysical J 85:1145–1164
17. Sippl M (1990) J Mol Biol 213:859–883
18. Holm L, Sander C (1993) J Mol Biol 233:123–138
19. Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD (2003) Nucl Acids Res 31:3497–3500. http://www.ebi.ac.uk/Tools/clustalw/
20. Chou KC, Shen HB (2007) Anal Biochem 370:1–16
21. Chou KC, Shen HB (2008) Nat Protoc 3:153–162
22. Panigrahi PR (2009) A Statistical Insight into Protein Structure. M. Sc. (Bioinformatics), CoE in Bioinformatics, Univ of Poona (Guide: Prof Joshi RR, IIT Bombay)
23. Patil RN (2010) Multivariate statistical analysis of some structural & functional features of proteins. MSc (Bioinformatics) CoE of Bioinformatics. Univ of Poona. (Guide: Prof Joshi RR, IIT Bombay)
24. Everitt BS, Dunn G (2001) Applied multivariate data analysis. Arnold, London
25. Hosmer DW, Lemeshow S (2000) Applied logistic regression, 2nd edn. Wiley, New York
26. Joshi RR (2007) Protein Peptide Letts 14:536–542
27. Fawcett T (2006) Pattern Recognit Lett 27:861–874
28. Hand DJ, Till RJ (2001) Machine Learning 45:171–186
29. Joshi RR, Jyothish NT (2010) Open Bioinformatics J 4:11–16