J Mol Model (2007) 13:275–282
DOI 10.1007/s00894-006-0141-z

ORIGINAL PAPER

Bayesian data mining of protein domains gives an efficient predictive algorithm and new insight

Rajani R. Joshi · Vivekanand V. Samant

Received: 31 March 2006 / Accepted: 28 July 2006 / Published online: 7 October 2006

Abstract Identification of structural domains in uncharacterized protein sequences is important in the prediction of protein tertiary folds and functional sites, and hence in designing biologically active molecules. We present a new predictive computational method of classifying a protein into single, two continuous or two discontinuous domains using Bayesian Data Mining. The algorithm requires only the primary sequence and computer-predicted secondary structure. It incorporates correlation patterns between certain 3-dimensional motifs and some local helical folds found conserved in the vicinity of protein domains with high statistical confidence. The prediction of domain-class by this computationally simple and fast method shows good accuracy of prediction—average accuracies 83.3% for single domain, 60% for two continuous and 65.7% for two discontinuous domain proteins. Experiments on the large validation sample show its performance to be significantly better than that of DGS and DomSSEA. Computations of Bayesian probabilities show important features in terms of correlation of certain conserved patterns of secondary folds and tertiary motifs and give new insight. Applications for improved accuracy of predicting domain boundary points relevant to protein structural and functional modeling are also highlighted.

Keywords Bayes' classification · Probability-computation · Proteins structural domains · PROMOTIF · PSIPRED

R. R. Joshi (✉) · V. V. Samant
Department of Mathematics,
Indian Institute of Technology Bombay,
Powai,
Mumbai 400 076, India
e-mail: rrj@math.iitb.ac.in

R. R. Joshi
School of Biosci. and Bioeng., IIT,
Bombay, India

Present address:
V. V. Samant
Persistent Systems, Pvt. Ltd.,
Pune 411 004, India

Introduction

Substructures of proteins that can fold independently are referred to as (structural) domains. Identification of such regions or domain boundaries of proteins is of crucial importance in structural biology and studies of structural genomics and functions of proteins [1, 2]. Computational methods of protein structure prediction and design require the information of the number of likely domains and their locations. This knowledge is also desired for accurate elucidation of NMR and crystallographic data to determine optimal tertiary structures.

Comparative sequence analysis by multiple sequence alignment and structural homology modeling has been used extensively for identifying the domains and tertiary structures of new proteins. However, in the absence of substantial sequential similarity or internal repeats, or, in the case of biased similarity, the method cannot be used or would risk giving biased/inaccurate predictions. Ab initio predictions or computational techniques independent of homology therefore provide promising alternatives. Notable among such methods for domain boundary predictions are—the DGS algorithm [3], SnapDRAGON [4], DomSSEA [5] with predicted secondary structure and PPRODO [6].

Domain Prediction by Guess (DGS) makes use of the narrow distribution of protein domain size on protein length. For proteins up to length 400 residues, it enumerates putative domain boundaries and calculates their relative likelihood under a probability model that considers only the size and segment number of predicted domains. The average accuracy (within ±20 residues around the 'true' boundaries) of its topmost predictions, for proteins up to length 400 residues, reported so far is almost 100% for single domains but 0% for two or more domain cases.

SnapDRAGON averages several hundred predictions obtained from ab initio simulations of the predicted 3-dimensional structures of a protein sequence to identify likely domain regions. The overall accuracy of this method for domain boundary prediction in a non-redundant sample of 185 single and 231 multi domain proteins was reported [4] to be about 72%. Its high computational complexity, however, limits its applicability.

Domain Secondary Structure Alignment (DomSSEA) is a fully automated method of domain assignment, using the alignment of predicted secondary structures of target sequences against observed secondary structures of chains with known domain boundaries as assigned by Class Architecture Topology Homology (CATH). The accuracy of prediction of domain numbers by this method for proteins of size up to 400 residues is reported [5] to be around 80% for single domain and 35% for two or more domain proteins.

PPRODO predicts domain boundaries of proteins from sequence information by a neural network. The network is trained and tested using the values obtained from the position-specific scoring matrix (PSSM) generated by PSI–BLAST. It is reported to predict the domain boundaries with an accuracy of about 66% using a resolution tolerance of ±20 residues for two continuous domain proteins of sizes up to 500 residues. However, it does not specify the discontinuous domains and it is computationally expensive.

Other prominent methods incorporate linker–region predictions between nearby domains; for example the *neural network* based program *DomCut*. [7] The automated method of Tanaka et al. [8] also uses neural networks trained on frequency data of single and multiple residue patterns present in linker segments. The prediction accuracies of these methods reported so far are only about 54 and 42%, respectively.

While the above methods focus mainly on prediction of domain boundaries and hence classifying as single, two or multi-domain classes, none predicts continuous and discontinuous domain categories before locating the likely domain or linker regions. Moreover, the accuracy of predictions of discontinuous domains[1] of the computational methods reported so far only ranges between 0 and 30%. Most of these methods begin with an estimate or random choice of the number of likely domains to locate the domain boundaries and refine them by some criterion derived from the representative (training) sets. Thus, there is a scope for improvement in the accuracy of prediction if the domain numbers and the type (continuous or discontinuous) are first identified with substantial accuracy.

We have developed a new ab initio method called Domain boundaries Prediction using Conserved local Patterns (DPCP) which first classifies the given protein sequence as *1d* (single domain), *2d* (two continuous domains) or *2dd* (two discontinuous domains) by a Bayesian machine learning technique. Using this classification, the method applies a heuristic algorithm to re-rank the DGS solutions and select the top five solutions accordingly. The heuristics are derived in terms of specific configurations of certain conserved tertiary structural patterns [9] and their statistical analogues in secondary structures that we found near the domain regions of the training sample. The domain boundary predictions using the latter show remarkable improvement [10] over DGS and the accuracy is also significantly better than DomSSEA when tested on a large validation sample of proteins of size up to 400 residues.

In this paper, we present the first phase (DPCP_0) of the method–prediction of domain numbers by a Bayesian machine-learning algorithm, which is derived from the new insight we obtained by modeling the probability distributions of certain conserved (geometrically invariant) patterns in the secondary and tertiary structures of the proteins in the training data set.

Materials and method: conserved structural patterns and Bayesian computations

Training and validation data set

The protein domain information like domain number and domain boundary positions were collected from the CATH (http://cathwww.biochem.ucl.ac.uk/latest/index.html) protein structure classification database [11] for a non-redundant set of protein chains with negligible (<1 to 15%) sequence homology. 2,150 proteins of sequence length (=number of amino acids) 70 to 400 were selected. Of these, 1,440 were in the single–domain class *1d*, 365 in *2d* and the remainder in *2dd*. Over 95% of the single–domain proteins here were also identified in single domain-class in the SCOP database [12] (web site: http://scop.mrc-lmb.cam.ac.uk/scop). This percentage was about 74–80% in the case of two–domain proteins; however, manual inspection showed CATH classification to be correct in 88 to 96% of the latter. The domain boundary points specified by CATH and SCOP

[1] Discontinuous Domains: At least one of the domains region is spread over disjoint portions of the protein sequence.

were within ±15 residues vicinity for two continuous domain cases. The same was true for at least one domain boundary in the case of two discontinuous domains. We have therefore chosen CATH as the reference for training and validation.

The primary sequences for these proteins were obtained from the Protein Data Bank (http://www.rcsb.org and [13]). Random subsets of about 30–40% of the proteins of each category were used in the *training sample* and the remaining in *validation sample* in the experiments described below. As the number of *1d* proteins in the *training sample* happens to be larger than the *2d* and *2dd*, only a subset (hereafter called *input set*) were used in estimating the probabilities for fitting the Bayesian model. This set has almost equal representations of all the domain classes under study; the subset of *1d* proteins contained in it was selected randomly. The minimum between-group variance technique was used to divide the *1d* training set for this selection. This ensures a statistically insignificant variation from one random subset to another. In view of the importance of protein length distribution, as shown by earlier studies [3, 5], the *input set* was partitioned according to the lengths of the proteins. As per the best statistical design, three length groups: 70–250, 251–300 and 301–400 residues were considered for data-analysis.

The entire training sample was used only for development of heuristics for predictive classification from the fitted Bayesian model. Jacknife type sampling technique is used in validation runs to make the validation sample larger in the case where the total number of proteins in a particular length group and domain-class is less than 70.

Search for standard 3D-motifs and correlation with 2D-folds

We first analyzed whether structures near the domain boundaries possess any characteristic tertiary patterns (motifs). Having found a statistically significant difference between the single versus multi-domain protein structures with respect to occurrence of some standard tertiary motifs, we investigated a Bayesian model that would compute the posterior probabilities of a domain-class given such a motif. However, as our main objective is to develop an efficient algorithm that uses only the primary sequence and a computationally-derived secondary structure to predict the domain-class for this sequence, we model these probabilities in terms of the corresponding secondary structural motifs using the estimates of necessary joint and conditional probabilities.

We have analyzed the frequency distributions of occurrences of standard secondary and tertiary structural motifs in the vicinity of domain boundaries in the proteins of the *input set*. The secondary structural motifs are specific configurations of the residue-wise secondary states predicted by PSIPRED [14] and the tertiary as predicted by PROMOTIF [9]. These motifs are hereafter referred to as 2D-motifs and 3D-motifs, respectively. Comparisons are also made with the occurrence of these patterns in single domain proteins.

The PSIPRED algorithm for secondary-structure prediction uses two feedforward neural networks, which perform an analysis on output obtained from PSI–BLAST (Position Specific Iterations of BLAST). The PSIPRED program (http://bioinfo.cs.ucl.ac.uk/psipred) predicts the secondary state helix (H), strand (E) coil (C) for each residue in the given protein sequence and also assigns a confidence level (on 0–9 scale) of each.

The PROMOTIF algorithm provides details of the location and types of structural motifs in proteins of known structure by analysis of Brookhaven format coordinate files. The program calculates the secondary structure of the protein using a variant of DSSP [15]. Further refinement of this raw data and identification of various super-secondary structures, and hence 3D-motifs from these, is done in a manner similar to that practised in crystallographic structure prediction using criteria based on consecutive bond length, *phi–psi* angles, distance between C_2-atoms of residues, orientation of side chains, etc. [9]. The PROMOTIF program (http://www.rubic.rdg.ac.uk./~gail/#Software) identifies presence of standard 3D-motifs, namely, helix, strands, hairpin loops, beta turns, gamma turns, and alpha-beta-alpha in different parts of the given protein's tertiary structure.

In order to analyze the distinction, if any, in the occurrence and/or the correlation between the 2D- and 3D-motifs near the domain boundaries against those in no-domain regions in the proteins of the *input set*, we have used a sliding-window approach. A window of size 40 residues is scanned along the protein sequences. The successive windows, W_1, W_2, etc, from amino to carboxyl termini of the sequence consist of segments of the first 40 residues, 41 to 80 residues, and so on. The last window has an overlap of 40-n consecutive residues with its previous in case the number of residues left for it is n where $n<40$. For two-domain proteins, additional window-segments consisting of 40 residues in the vicinity (±20 residues) of the domain boundaries of all the continuous and discontinuous domains are used.

For successive windows of each protein in the *input set*, our program searches for (I) the presence of the standard 3D-motifs as identified by PROMOTIF and (II) the occurrence of specific 2D-motifs as consecutive patches of H, E and C predicted by PSIPRED. Simultaneously with these, the frequency distribution of occurrence of domain boundary points (*dbp*) in the successive sliding windows is also obtained.

Table 1 Joint probabilities of the 3D-motifs and same type of 2D-motifs: these are estimated as average relative frequencies in the *input set*

Motif type	Two continuous domain proteins		Single domain proteins	
	Pr{same type of 2D- & 3D-motif in W^*}	Pr{same type of 2D- & 3D- motif *at same location* in W^*}	Pr{same type of 2D- & 3D-motif in W^*}	Pr{same type of 2D- & 3D- motif *at same location* in W^*}
HLX	0.858696	0.152174	0.951673	0.217472
STR	0.997382	0.108639	0.905000	0.146538
LOOP	0.384518	0.178396	0.348900	0.247804

Here W^* denotes the window(s) under consideration. Sample results for input proteins in length-group 70–250 are shown here.

Domain boundary point locations and 3D-motif distributions

The frequency plots of shorter proteins (length 70 to 250) showed that the sequence–portions covered by the windows W_2 to W_4 have a high propensity to contain domain boundary points (dbp). For longer proteins (length 251 to 400) the portions covered by W_3 to W_7 are found to be potential dbp regions. We shall denote these window unions as W^*:

$$W^* = W_2 \cup W_3 \cup W_4;$$

for proteins in length group 70 to 250 and,

$$W^* = W_3 \cup W_4 \cup W_5 \cup W_6 \cup W_7; \quad (1)$$

for longer proteins.

The motif frequencies C_i; i=1,....., N were calculated for each 3D-motif type M_j; j=1, ..., 5, for each protein in the input data set; where N is the number of windows suitable for the protein under consideration; M_1 to M_5, respectively, imply helix, strands, beta-turns/gamma-turns, hairpin loops and alpha-beta-alpha. On the basis of these frequency values, the following three types of motif frequency plots are computed for each motif type:

f_0: Number of proteins for which C_i (M_j)=0
f_1: Number of proteins for which C_i (M_j)=1
f_{1+}: Number of proteins for which C_i (M_j)≥1

The frequency-plots of f_1 and f_{1+} for M_1 (hereafter denoted by HLX), M_2 (STR) and M_3 (LOOP) and that of f_0 for M_2 are found significantly different in the case of single (*1d*) and multi-domain (*2d* or *2dd*) proteins. Therefore, for fitting the Bayesian model, only the events that correspond to these motifs are considered over the prominent windows—i.e. the segments with a high propensity of containing a *dbp*. The necessary details are outlined in the Bayesian computations: prediction of protein domain class and Results and discussion.

Estimates for joint probabilities of 2D- and 3D- motifs

Frequency distributions of standard 3D-motifs, namely, HLX, STR, and LOOP, predicted by PROMOTIF are obtained for the protein sequences in the *input set*. Based on frequency analysis in the prominent windows, the following configurations of the PSIPRED output are found most suitable to define corresponding 2D-motifs: A consecutive patch of six or more Hs predicted by PSIPRED with a confidence level ≥8 is regarded as the 2D-motif of type HLX; a consecutive patch of two or more Es with confidence level ≥5 as 2D-motif of type STR; and four or more consecutive Cs with a confidence level ≥6 as the 2D-motif of type LOOP.

The marginal, conditional and joint probabilities of occurrence of the 3D- and corresponding 2D-motifs in the prominent window(s) are then estimated for the *input set*. Some of these estimates are shown in Table 1. These probabilities are also computed for overlapping segments of a fixed size within window cluster W^*; this is done mainly to explore the nature of these probabilities in the regions closer and farther away from the dbp, if any, in the window of interest.

Optimal location G_X

The optimal combination of overlapping segments in W^* is identified as the one for which there is maximum difference between the estimated joint probabilities for the single and two domain proteins of same length group in the input set. We denote it by G_X for motif type X. The best results (in terms of distinction between different domain classes) for the *input set* show the optimal segment size for the different motif types as follows: segment size=25 residues for HLX, segment size=15 residues for STR and LOOP. Overlap of five residues between consecutive segments is used for each case.

Bayesian computations: prediction of protein domain class

Using the probability distributions estimated for the input set, a Bayesian algorithm is trained to classify a protein sequence as *1d*, *2d*, or *2dd*.

Step 1: Empirical Bayesian model for 2D-motifs

a) Prior Probabilities of domain-classes:

p_k^0=Prior probability of {domain numbers=k}; k=1, 2; is computed as,

$$= n_k/n \quad (2)$$

Where, n=cardinality of the *input set* and n_k=no. proteins (in this set) with k domains.

For example, for proteins in the length group 70–250, this probability is found to be approximately equal to 0.84 if k=1 and 0.16 if k=2. This is in accordance with the lengthwise domain–number distribution reported by [3]. Similar is the case for the prior probabilities for longer proteins.

b) Conditional probabilities of events defined for 2D-motif distribution:

The following probabilistic events in W^* are found to be most informative for distinguishing between single and two–domain proteins using the 2D-motifs

E_1: No. of HLX=1 E_3: No. of STR=1 E_5: No. of LOOP=1
E_2: No. of HLX>1 E_4: No. of STR>1 E_6: No. of LOOP >1

c) Posterior Probabilities are computed using Bayes' Theorem as:

$$P(k|E_i) = \frac{p(E_i|k)p_k^0}{\sum_{k=1}^{2} p(E_i|k)p_k^0} \quad (3)$$

Where k denotes the number of domains. The prior probabilities p_k^0 are given by Eq. 2 and the conditional probabilities $p(E_i|k)$ are estimated using the relative frequency plots from the *input set*.

Step 2: Computation of Probabilities in terms of 3D-motifs:

Noting that the denominator of Eq. 3 is common for all k, we consider only the terms associated with the numerator for computation of the decision functions for Bayesian classification.

Let D_i, i=1, ..., 6, denote the analogues of events E_1 to E_6 for the corresponding 3D-motifs.

$$\psi^*(k|D_i) = P(k|E_i)\lambda(E_i, D_i, |k) \quad (4)$$

Where $P(k\,|\,E)$ denotes the posterior probability given by Eq. 3 above; and λ() is estimated in terms of the joint probabilities (e.g. see Table 1) of the corresponding 2D- and 3D-motifs in the specified region.

Three types of experiments are carried out considering different portions in W^*. *Experiment I*: entire portion W^* of interest; *Experiment II*: the exact location in W^* where the 2D- and 3D- motifs of same type would occur; *Experiment III*: the optimal combination G_X of overlapping segments in W^* with respect to the specific motif type say X; 'optimality' and G_X as defined in Estimates for joint probabilities of 2D- and 3D- motifs.

The third set of experiments is found to give best results in terms of distinct plots of ψ * for single and multi-domain proteins in the training sample. Sample outputs of some results in this regard are shown in Fig. 1a–d.

Extension to classifying single, two-continuous, and two-discontinuous domains

The same method is extended for classifying single, two-continuous and two-discontinuous-domain proteins. The protocol described above is used for the estimations of respective prior, conditional and posterior probabilities and the decision function ψ^*; here k is assigned three distinct integer labels for the class-types *1d*, *2d* and *2dd*. In this case also, the results showed the third experiment, viz.,

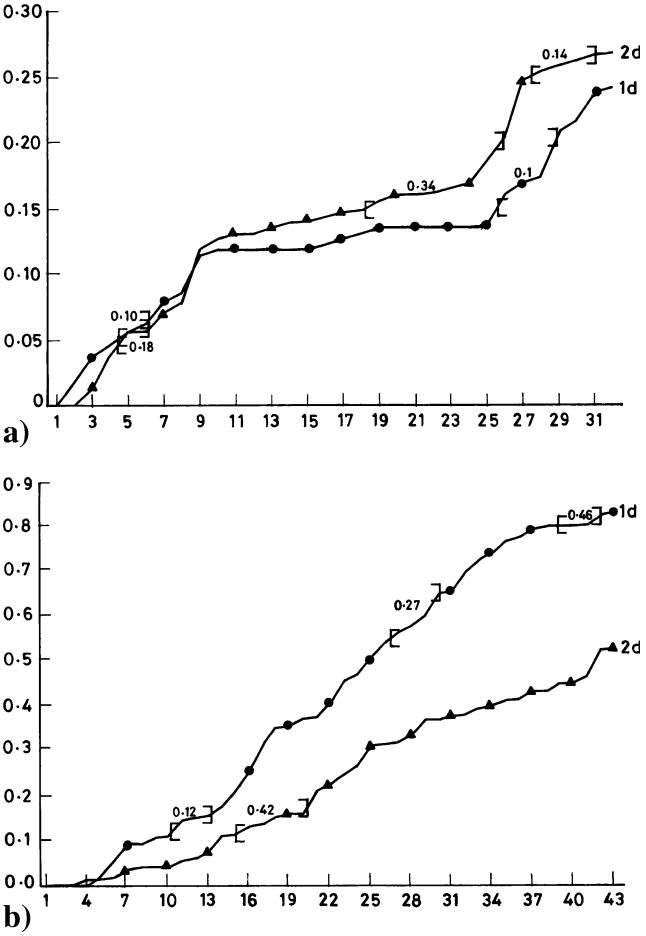

Fig. 1 *Decision Function* plots ψ^* (*class*|*event*) for Experiment III where *event* is defined in terms of the *3d-motif*: **a** protein length-group 70–250; *class 2d*; *event* HLX ≥1; **b** protein length-group 251–300; *class 1d*; *event* HLX ≥1; **c** protein length-group 301–400; *class 2d*; *event* STR ≥1; **d** protein length-group 251–300; *class 2dd*; *event* LOOP ≥1. Other details as explained in the text (see Illustrations in Unexpected likelihood of motifs—new insights)

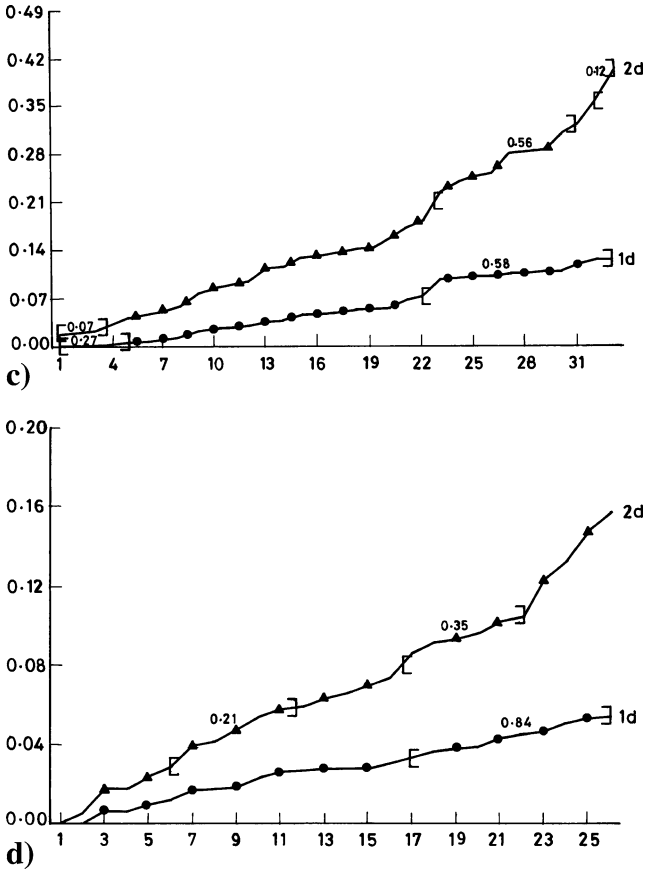

Fig. 1 (continued)

computing the desired probabilities in the optimal segments of W^* to be the best for predictions.

Prediction of domain-class for validation sample

For a protein in the validation data set, we use only its primary sequence and the PSIPRED predicted secondary structure. Identifying its length group and the events E_i present in the corresponding optimal segments of W^*, the decision function ψ^* is computed as per (4) with k representing distinct integer labels for the class-types *1d, 2d* and *2dd*.

In the conventional Bayesian sense, the class for which the decision function is maximum would be selected. However this is not strictly applicable here, as different proteins in the same domain-class may have different ψ^* with respect to the same event because their optimal segments could be different. Moreover, the conventional approach may be inappropriate owing to the possibility that, for a protein sequence, approximately the same 'maximum value' of ψ^* is obtained for two different domain-classes.

We therefore use certain data-driven heuristics [16, 17] for classification.

Heuristics for predictive classification

During trials on the training sample, the distinct distribution of ψ^* in different classes with respect to common events gave rise to heuristics like—"If the estimated decision function ψ^* with respect to certain event lies in a particular interval, then a particular type of domain class..."; Or, "If an event is absent then the odds ratio in favor of domain type, say k_1 against type k_2 are $\beta 1{:}\beta 2$"; Or, combinations of these kinds of heuristics.

Based on its performance for the training sample, each heuristic is also assigned a worth, a quantity directly proportional to its overall success rate of correct predictions. The heuristics with higher worth are given preference in case of a tie while predicting the domain-class of a protein sequence in the validation sample. Suppose there is also a tie on the worth, then the class that has higher prior probability would be selected.

Results and discussion

Proteins in the validation sample are treated as uncharacterized, i.e., no knowledge other than their primary sequences and PSIPRED predicted secondary structures, is used. Predictions of their domain-class are made using the Bayesian heuristic algorithm described above. The predicted classes are verified against the domain-class allocated by the CATH database. Accuracy is estimated as the percentage of proteins where the predicted class and the CATH-assigned class are the same. The average prediction accuracy for single (*1d*), two-continuous (*2d*) and two-discontinuous (*2dd*) domain proteins in the validation samples is found to be 83.3, 60 and 65.7%, respectively.

The performance of our method was compared with that of the best-known computational methods DGS and DomSSEA on the same validation sample. The predictions of DGS and DomSSEA are made using their web utilities [3] (as of Oct.–Nov. 2005) and http://bioinf.cs.ucl.ac.uk/dompred (latest version). The accuracies of DGS for class *1d*, *2d* and *2dd* are: 100, 0 and 0%, respectively. The corresponding accuracies of DomSSEA are: 72, 24 and 0%.

The accuracies and false-prediction details of our method (DPCP_0) when applied are shown in Table 2 below for different sequence-length groups.

Bayesian learning models have been used extensively in AI systems with remarkable success. Their power lies in computational simplicity and sequential updating of acquired knowledge without any constraint of specific probability distribution. Another important factor is the flexibility to incorporate, in a modular way, empirical probabilistic rules for knowledge acquisition. This is why Bayesian data mining continues to be most suitable for

Table 2 Percentages of correct, false–positive and missed classifications

Protein Sequence Length	By CATH	% of proteins (of CATH given domain-class) predicted in domain-class		
		1d	*2d*	*2dd*
70–250	*1d*	**83.50**	6.59	9.91
	2d	29.60	**57.40**	13.00
	2dd	31.77	1.23	**67.00**
251–300	*1d*	**85.14**	14.86	0.00
	2d	23.21	**64.30**	12.49
	2dd	18.18	11.82	**70.00**
301–400	*1d*	**81.12**	15.03	3.85
	2d	35.59	**58.0**	6.41
	2dd	26.67	13.33	**60.00**

The accuracies (correct prediction percentages) are shown in boldface.

unstructured, uncertain and uncharacterized data with high complexity, heterogeneity and multiplicity of information/knowledge contained in them. The protein structural data are the best example of this kind.

We have used this approach for 'extracting' the nonlinear random features that play a key role in characterizing single or multiple-domain tertiary structures. In particular, our Bayesian machine-learning algorithm is found to be efficient and most accurate (as compared to the best known computationally simple methods reported so far) in predicting whether an uncharacterized protein sequence would have *1d*, *2d*, or *2dd*.

Despite the fact that optimal segments, and hence the value of the decision function ψ^* for the events associated with these segments would be different for each protein, we have found the variation of ψ^* to be quite narrowly distributed within a common length-group and domain class. Modeling the posterior probability functions with respect to these factors using the ψ^* curves for different length groups in our computational experiments would be an interesting theoretical research problem with promising scope in structural biology, proteomics and evolutionary studies on proteins.

Scope for improvement of the DBP prediction

Accurate prediction of the domain class promises improvement in the domain-boundary prediction by the computational methods that begin with a random guess or a guess based on protein-sequence length. Our method DPCP that begins with the (predicted) knowledge of the domain class and then computes expected domain-boundary points is an example. The details of this method have been reported separately [10]. The accuracy of its top-ranked solutions is found to be around 84, 64, 58 and 53% for predicting the dbp of two continuous domain proteins in the length groups 70–250, 251–300, 301–350 and 351–400, respectively. This is either better or closely comparable with the best-known methods so far. Its performance is significantly better than the latter in predicting the dbp of *2dd* proteins; for the corresponding length groups the prediction accuracies in this case are around 74, 80, 62 and 60%.

Unexpected likelihood of motifs—new insights

Computational methods of structural biology often focus on sequential and structural homology with some reference set. The unique aspect of our Bayesian approach is that it does not require homology with any reference set. This ab initio method exploits the random variation as well the similarity of certain structural features within a domain-class and the distinct statistical nature of these between different domain classes. The prediction results show good potential of this method.

It is important to note that, although the sequences in our data sets had no homology and the best results are found for the computational experiments on optimal segments, which could be different for different proteins, the ψ^* values for proteins in the common length group and domain class are clustered in some small intervals only—the within cluster variance (of computed values of ψ^*) is of the order of 0.003. Interestingly, these intervals are mostly distinct for the single and two-domain cases; hence the Bayesian heuristics have worked efficiently. The heuristics pertaining to ψ^* for any of E_1, E_2, E_3 falling in specific intervals are found to make correct predictions with 68% to 92% accuracy.

Each curve in Fig. 1 is plotted only for distinct values of ψ^* (*w.r.t.* the particular event and domain class in a given protein length group). The estimates of ψ^* for different proteins in the input sample are found to cluster around of these distinct values. Some of these clusters and the confidence levels of predicted class (as *1d* or *2d* or *2dd*) when ψ^* belongs to the corresponding intervals are also shown along the curves. Thus, the Figures comprehensively illustrate the trend, pattern and extent of variation in posterior probability and also provide qualitative as well as qualitative comparison between the classes of interest.

Illustration

In each Figure, the integers *1*, *2*, ..., *m* along the horizontal axis are only indicators of the *m* distinct values of, ψ^*; the ψ^* values are represented along the vertical axis. The ψ^* values between a pair of two consecutive points, marked by "[" and "]", respectively, on each curve represent an interval used in some heuristic to predict the corresponding domain class distinctly. The (relative) likelihood of this prediction being correct is shown (on a 0 to 1 scale) on the curve between that pair of marked points. If there are no

marks shown for corresponding interval of values on the other curve, or if the interval is not attained on this curve, then the chance of the class represented by the latter is 0. For example, the two topmost pair of marks in Fig. 1b show that if the computed ψ* lies in the interval [0.55, 0.65] or [0.75, 0.85], then predict single domain (*1d*) class. The likelihood of this decision being correct is 0.27+0.46 (=0.73), whereas that of predicting the two–continuous domain (*2d*) class is 0.

For each length-group, in general, the heuristics found most useful in classifying between *1d* and *2d* are (I) those defined in terms of non-occurrence of the events of STR and/or HLX and (II) those defined for different intervals of ψ* associated with these events. LOOPs did not play any significant role here. On the contrary, the heuristics associated with occurrence or non-occurrence of LOOP are more applicable while distinguishing the *1d* and *2dd* classes (Fig. 1d for example). These kinds of observations, together with the fact that linker regions between domains often play significant roles in the activity and flexibility of the proteins, indicate the possibility of nature favoring certain geometries for certain functions in the protein repertoire.

Further, the intermediate results of data mining for our Bayesian machine-learning algorithm elucidate some unexpected features with respect to the occurrence of structural motifs.

For example, though secondary (local) helical patterns are generally expected to occur in the vicinity of the *dbps* (e.g. [18, 19]), the occurrence of their statistical analogues in 3D-motifs does not seem to be significantly different in the non-domain region. For example, the behavior of the variation in their likelihood (ψ*) is either mostly inconclusive (e.g. Fig. 1a) for *2d* proteins or is more prominent (e.g. Fig. 1b) in the class of *1d* proteins. On the other hand, several heuristics derived in terms of ψ* for given STR are found to be significantly useful in predicting the *2d* class because of the distinct shapes and trends of the curve for this class and that for the *1d* class (e.g. Fig. 1c). The consistent dominance (higher ψ*) of the curve representing this class is also notable.

The heuristic associated with the absence of HLX is found to have an odds ratio in favor of the *1d* vs *2d* of 34:62, whereas this ratio for the heuristic dealing with absence of STR is 51:36 for *1d* vs *2d* and 35:6 in the case of *1d* vs *2dd* class.

The Bayesian probabilistic approach used here could also be extended with incorporation of certain sub-motifs, geometrically invariant patterns and clusters of amino acids [20], to study the evolution of the protein universe and associated aspects of protein structural and functional genomics. Because of the inherent heterogeneity of the protein data, new statistical experimental designs and non parametric multivariate analysis would be required to extend the approach to longer protein chains and multiple domain classes.

Acknowledgement This work is part of an R & D project undertaken by the first author. The author is thankful to the Department of Biotechnology, Govt. of India, for the financial support.

References

1. Joshi RR, Jyothi S (2003) Comput Biol Chem 27:241–252
2. Jyothi S, Mustafi SM, Chary KVR, Joshi RR (2005) J Mol Mod 11:481–488
3. Wheelan SJ, Marchler-Baucer A, Bryant SH (2000) Bioinformatics 16:613–618
4. George RA, Haringa J (2002) J Mol Biol 316:839–851
5. Marsden RL, McGuffin LJ, Jones DT (2002) Protein Sci 11:2814–2824
6. Sim J, Kim SY, Lee J (2005) Proteins Struct Func Genet 59:627–632
7. Suyama M, Ohara O (2003) Bioinformatics 19:673–674
8. Tanaka T, Kuroda Y, Yokoyama S (2003) J Struct Func Genomics 4:79–85
9. Hutchinson EG, Thornton JM (1996) Protein Sci 5:212–220
10. Joshi RR, Samant VV (2006) J Mol Model (in press)
11. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM (1997) Structure 5:1093–1108
12. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) J Mol Biol 247:536–540
13. Berman HM, Westbrook J, Feng Z, Gillilad G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) Nucl Acids Res 28:235–242
14. McGuffin LJ, Jones DT (2000) Bioinformatics 16:404–405
15. Kabsch W, Sander C (1983) Biopolymers 22:2577–2637
16. Berthold M, Hand DJ (eds) (1999) Intelligent data analysis. Springer, Berlin Heidelberg New York
17. Baldi P, Brunak S (2001) Bioinformatics: the machine learning approach (adaptive computation and machine learning), 2nd edn. The MIT Press
18. Sowdhamini R, Rufino SD, Blundell TL (1996) Fold Des 1: 209–220
19. Batencourt MR, Skolnick J (2001) Biopolymers 59:305–309
20. Stanfel LE (1996) J Theoret Biol 183:195–205