J Mol Model (2006) 12: 943–952
DOI 10.1007/s00894-006-0116-0

ORIGINAL PAPER

Rajani R. Joshi · Vivekanand V. Samant

Fast prediction of protein domain boundaries using conserved local patterns

Received: 30 December 2005 / Accepted: 9 March 2006 / Published online: 29 April 2006

Abstract We have found certain conserved motifs and secondary structural patterns present in the vicinity of interior domain boundary points (dbps) by a data-driven approach without any a priori constraint on the type and number of such features, and without any requirement of sequence homology. We have used these motifs and patterns to rerank the solutions obtained by the well-known domain guess by size (DGS) algorithm. We predict, overall, five solutions. The average accuracy of overall (i.e., top five) predictions by our method [domain boundary prediction using conserved patterns (DPCP)] has improved the average accuracy of the top five solutions of DGS from 71.74 to 82.88 %, in the case of two-continuous-domain proteins, and from 21.38 to 80.56 %, for two-discontinuous-domain proteins. Considering only the top solution, the gains in accuracy are from 0 to 72.74 % for two-continuous-domain proteins with chain lengths up to 300 residues, and from 0 to 62.85 % for those with up to 400 residues. In the case of discontinuous domains, *top_min* solutions (the minimum number of solutions required for predicting all dbps of a protein) of DPCP improve the average accuracy of DGS prediction from 12.5 to 76.3 % in proteins with chain lengths up to 300 residues, and from 13.33 to 70.84 % for proteins with up to 400 residues. In our validation experiments, the performance of DPCP was also found to be superior to that of domain identification from secondary structure element alignment (DomSSEA), the best method reported so far for efficient prediction of domain boundaries using predicted secondary structure. The average accuracies of the topmost solution of DomSSEA are 61 and 52 % for proteins with up to 300 residues and 400, respectively, in the case of continuous domains; the corresponding accuracies for the discontinuous case are 28 and 21 %.

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

Keywords Protein structures · Protein domain boundary points · Nonparametric statistics · DGS · PSIPRED

Introduction

The characterization and prediction of the linker regions and protein domain boundaries are important in tertiary-structure prediction from the primary sequence of proteins and studies on their structural and functional genomics. Several computational methods have been reported recently that employ (1) the findings of exhaustive comparative sequence and linker-structure analyses of single and multidomain proteins [1], and (2) the narrow distribution of domain boundary locations with respect to the lengths of proteins. [2].

While homology-based methods extensively search for specific patterns, e.g., regions of low-complexity, transmembrane segments, coiled-coils, or long stretches of repeated residues, particularly proline, glutamine, serine, or threonine, to model linker regions or the separation of globular domains, some ab initio methods have also attempted the identification of such patterns for linker-region prediction. Notable among these methods is the *neural network*-based program DomCut [3]. The automated method of Tanaka et al. [4] also uses neural networks trained on frequency data of single and multiple-residue patterns present in linker segments. However, the overall prediction accuracies of these two methods are only about 54 and 42 %, respectively.

Earlier studies on protein-structure databases have shown that the size of protein domains is not fixed, and also that all domain lengths are not equally likely to occur. It has been observed that protein domain lengths follow a narrow distribution and, furthermore, that domains identified in the three-dimensional-structure database most often contain a single chain-continuous segment [5, 6]. In view of the statistical significance of protein length in the location of domain boundary points (dbps), the domain guess by size (DGS) program enumerates putative domain boundaries. Using the length distribution in a large training

sample, DGS assigns log-likelihood scores to dbps at residue numbers 20, 40, etc. [2]. The solutions are ranked according to the log-likelihood score. The output lists the top ten solutions. As the majority of proteins up to length 400 are found to be single-domain, the first-ranked solution of DGS is mostly *L*, which is equal to the multiple of 20 nearest to the N terminus.

The domain identification from secondary structure element alignment (DomSSEA) algorithm [7] offers delineation of continuous domains, fully automatic 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 reported average accuracy of the top prediction (with a resolution of ±20) by DGS is around 77 % for single-domain proteins and 30 % on average for structured multidomain proteins with up to 400 residues. The top prediction average success rate of DomSSEA in correctly assigning a domain number to the representative chain set is 73.3 %. However, the top prediction for location of domain boundaries is reported to be correct for only about 24 % of the multidomain set. The program Prediction of Protein Domain Boundaries (PPRODO) [8] predicts domain boundaries of two-continuous-domain proteins using a neural network trained and tested by the values obtained from the position-specific scoring matrix generated by the Position-Specific Iterated Basic Local Alignment Search Tool; the overall accuracy of classification between single- and two-continuous-domain proteins with up to 500 residues is reported to be around 66 %.

Our method [domain boundary prediction using conserved patterns (DPCP)] applies a knowledge-based and Bayesian approach to test whether a protein has single or multiple domains using the primary sequence and some maps of geometrically invariant patterns in the predicted secondary structure. For predicted two-domain chains, it then computes the expected dbps by reranking the DGS solutions. This refinement is done using certain heuristics on the predicted secondary structure of the given protein sequence. This second part of DPCP, computation of likely domain boundaries in the two-domain class, is the focus of the present paper. Throughout this paper, the words domain boundary point refer to the interior dbps, i.e., the domain end-points other than the N and C termini of the protein chains.

Data and algorithm

The protein-domain information, such as domain number and domain boundary positions, were collected from the CATH (http://cathwww.biochem.ucl.ac.uk/latest/index.html) protein structure classification database for a nonredundant set of protein chains with negligible (<1 to 12 %) sequence homology. A total of about 2,140 proteins of sequence lengths (number of amino acids) 70 to 400 were selected. Of these, 1,440 had single domains, 365 had two continuous domains, and the remaining had two discontinuous domains. Over 95 % of the single-domain proteins here were also identified in the single-domain class in the Structural Classification Of Proteins (SCOP) database (http://scop.mrc-lmb.cam.ac.uk/scop). This percentage was about 80 % in the case of the two-domain proteins; however, manual inspection showed the CATH classification to be correct in 88 to 96 % of the remaining set. The dbps predicted by CATH and SCOP were within ±15 resolution for the 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 the CATH predictions as the reference data for training and validation.

The primary sequences for these proteins were obtained from the Protein Data Bank (http://www.rcsb.org/pdb/). Random subsets of about 40–50 % of the proteins in each category were used as *training samples* and the remainder as validation samples in the experiments described below. The jackknife-type sampling technique was also used in validation runs to make the validation sample larger in the cases where the total number of proteins was less than a hundred.

DPCP_0: predictive classification of single- and two-domain proteins

We had analyzed a random set of about 500 two-domain proteins (including those in the *training sample* of DPCP) to study the structural conservations in the vicinity (±20 residues) of protein domain boundaries. The algorithm of Tendulkar et al. [9] was used to locate the presence of geometrically invariant tertiary-structure patterns [10] of *octapeptides* in these segments. For this, the tertiary folds of a sliding window of size-8 residues are mapped onto geometrically invariant descriptors, such as three-dimensional graphs. These residues are then clustered based on structural similarity. Only the similarity of geometrical shape is sought here—independent of the amino acid sequences of the *octapeptides*. In our experiments, *helical motifs* were found to be predominant in clustering the two-domain proteins by this technique. This motivated us to use the secondary folds correlated with these and other standard tertiary motifs to distinguish between single- and two-domain proteins.

Frequency distributions and correlation maps of the three-dimensional standard motifs, *helix*, *strands*, *hairpin loops*, *beta turns*, *gamma turns*, *alpha–beta–alpha*, etc., identified by PROMOTIF (http://www.rubic.rdg.ac.uk./~gail/#Software) were analyzed against those of the Protein Structure Prediction Server (PSIPRED) (http://bioinf.cs.ucl.ac.uk/psipred/) predicted local folds of *helix*, *strands*, and *loops* in the secondary structures. Further data mining using a sliding window of 40 residues on the primary sequence showed statistically significant differences in joint occurrences of the three-dimensional and secondary motifs in the single- vs multidomain proteins. Posterior probabilities of the single- and two-domain classes were

computed using a nonparametric Bayesian model for these data. Prediction of single domains (*1d*), two continuous domains (*2d*), or two discontinuous domains (*2dd*) was then made using the criteria of consensus Bayesian classification for statistically significant secondary (local) patterns in a given window, or posterior probabilities of the chosen class above a certain threshold. The major steps of this algorithm are shown in Appendix 1. Details of the frequency plots and derivation of the Bayesian decision functions are given in our separate paper (Joshi and Samant, personal communication), along with computational experiments and results.

Testing on validation samples from the CATH and SCOP data banks gave 83.4 % correct predictions in the *1d* class, 60.2 % in the *2d* class, and 65.5 % for the *2dd* class; whereas these accuracy measures of DGS, implemented on the same samples, are 76.7, 0, and 0 %, respectively. The accuracies of DomSSEA (http://bioinf.cs.ucl.ac.uk/dompred/) for these classes for the same validation samples are found to be 55, 68.1, and 0 %, respectively.

DPCP—domain boundary prediction

For a given protein predicted to have two continuous or two discontinuous domains, *initial solutions* to the dbps are obtained as DGS solutions ranked 2 to 10. (The top solution of DGS is not used, as it usually corresponds to the single-domain case).

The secondary structure of the protein chain under consideration was obtained using PSIPRED, which is known to be the most accurate software for this purpose at present. It predicts the secondary state of each residue as H (helix), C (coil), or E (strand) and also assigns a reliability (of prediction) measure to each as an integer between 0 and 9.

Data mining for conserved patterns

Considering the possibility of modifying DGS predictions by using additional information, we have explored the characterization of secondary-structure patterns, if any, in the vicinity of the dbps. Thorough data mining of the training sample using the geometrically invariant local patterns has shown *helical* patches of six or more residues, predicted by PSIPRED with a reliability coefficient ≥6, as statistically significant in the neighborhood (±15 residues) of the dbp. Such helical patches are referred as *conserved patterns* hereafter.

Several experiments were conducted on different subsets of the training sample to minimize variance in the location of such patterns within and outside the vicinity of the dbps. It may be noted that, because of the highly heterogeneous, aperiodic, and random nature of the protein structural data, no single probability distribution can be fitted in the conventional statistical sense. We had only used the relative frequency of occurrence of conserved patterns of different lengths in different parts of the protein chains in the training sample to assess the possibility of a dbp within or around such patterns. Preference scores for locations of the conserved pattern relative to that of the dbp were assigned as directly proportional to the relative frequencies estimated from the training sample. These scores are used as the heuristic guideline in the case of a tie, while labeling the initial solutions as described in Phase1 below.

Table 1 A sample output of the labeling algorithm for the protein 1qag

Dbp	Predicted secondary structure of segment (dbp −15 to dbp +15). Residue-wise confidence levels of prediction assigned by PSIPRED	Labels
140	**HHHHHHHHHH**CCCCCCCEEEECCHHHHHHHHH 9 9 9 9 9 9 9 8 6 1 56 67 7 1 1 4 4 4 4 2 0 0 1 0 1 0 4 7 9	F
100	HHCCCH**HHHHHHHHHHHHHHHHHHHHHHHHH**HH 1 2 37 5 5 7 8 9 9 9 9 9 9 9 9 9 8 8 8 8 8 8 8 8 7 6 5 2	A
120	HHHHHCHHHHHCCCCCC**HHHHHHHHHHHHH**CC 8 8 8 8 8 8 8 7 6 5 2 0 2 3 6 7 6 7 8 9 9 9 9 9 9 9 9 9 8 6 1 5	C
160	CCHHHHHHHH**HHHHHHHHHH**HHHCCCHHHCCCC 4 2 0 0 1 0 1 0 4 7 9 9 9 9 9 9 8 8 7 4 4 1 0 1 4 1 2 1 3 7 5 7	E
80	**HHHHHHHHH**CCCCCCCCCCCHHHHCCCHHHHHH 9 9 9 9 9 9 9 7 7 9 8 2 16 8 9 7 0 0 2 1 2 3 7 5 5 7 8 9 9 9	G
180	HCCCHHHCCCCC**HHHHHHHHHHHHHHHHH**CCCC 1 0 1 4 1 2 1 3 7 5 7 87 8 8 9 9 9 9 9 9 9 9 9 9 8 6 1 9 8 3	B
60	CCCCCCCCCCCCCCCHHHC**HHHHHHHHHHH**CCC 7 8 7 5 77 6 7 7 7 6 5 2 0 0 3 0 57 8 9 9 9 9 9 9 9 7 7 9 8	D
200	HHHHHHHCCCCCCCCHHHHCCCCCCHHHHHH 9 9 9 9 9 8 6 1 9 8 3 0 0 3 6 3 3 3 2 1 5 8 9 8 2 8 8 9 9 9 9	
40	CHHHHHHHHHHHHHHHHHHHCCCCCCCCCCCC 0 6 8 8 8 8 7 5 9 9 9 9 9 9 9 9 9 8 7 1 7 8 7 5 7 7 6 7 7 7 6	

Sample output of the DPCP_Phase1 for the protein 1qag (chain A, primary length 226) is illustrated here. The "*conserved*" helical patches described in the section "Data mining for conserved patterns" are presented in bold. The solutions (dbp) are those predicted by DGS; these are shown here in the order of the ranks (second to ninth) given by DGS. The last two solutions are not considered for labeling, as they fall in the masked portion near the two termini of the protein chain

DPCP_Phase1: labeling of DGS solutions

Using the predicted secondary structure, segments of ±15 residues around the initial solutions for dbps are scanned for the presence of the conserved patterns defined above. Only those initial solutions that are outside the masked portion, i.e., those which are away from 20 % of the portion of the protein length from the N and C termini, are selected.

The initial solution (for the dbp) corresponding to the segment which contains the longest patch of the conserved pattern is labeled *A*; the initial solution corresponding to the segment having the second longest patch is labeled *B*, and so on. In the case of a tie with respect to this criterion, the higher preference score of the location of the patch is given priority. As for the heuristic guideline, the segment in which the patch contains the corresponding initial solution of the dbp is given the highest priority in this case. Successively lower priorities are given to the segments where the conserved patterns lie on the left and right of the corresponding initial solutions. If there still is a tie in the ordering of the segments, the higher total reliability coefficient attached (by PSIPRED) to the conserved pattern is given greater priority.

This labeling procedure is illustrated in Table 1. In this table, the solution (dbp=100) in the second row is labeled *A* because the secondary structure of the segment around it contains the longest patch of the conserved pattern. The length of this patch happens to be the same (9) for two solutions, namely, dbp=140 and dbp=160. These are labeled *F* and *E*, respectively, because the latter (residue number 160) is present inside the marked patch (the segment around residue number 160 consists of residues at primary chain positions 145 to 175).

Irrespective of what label is assigned to the second-ranked DGS solution by the above procedure, this solution is also labeled *I*.

In our training experiments, the solutions labeled *A* to *D* and *I* are found to include the best choices in terms of accuracy of predicting the domain boundaries in the training sample. These solutions are ranked among themselves, as described in the next section.

DPCP_Phase2: ranking of domain boundary predictions labeled A, B, C, D, and I

Testing the labeled solutions showed that the alphabetic order of the labels would not be a suitable choice for ranking the solutions. Heuristics derived from nonparametric statistical data-mining experiments on *2d* and *2dd* proteins in the training sample were therefore used to rank the five predictions. The best statistical design corresponding to the maximum accuracy of these heuristics suggested separate consideration of different groups in terms of the length of the protein chain. The approach is explained in Appendix 2.

Ranking of predicted domain boundaries for length group of 80–250 residues

The following heuristics were applied for ranking the predictions for proteins with 80 to 250 residues.

if (protein length<=240) {

Out of the two predictions *A* and *B*, choose the one that is closer to the midpoint of the protein length.}else {

Out of the two predictions *A* and *B*, choose the one that is far from the midpoint of the protein length.}

Out of the selected solution (*A* or *B*) and *I*, the one closer to the midpoint of the protein chain is ranked as the top and the other as the second. The third rank is assigned to whichever prediction, of *A* and *B*, was not selected by the above criteria.

If both the predictions are equidistant from the midpoint of the protein length, then choose the prediction labeled *I* as the top rank, and the second and third ranks are assigned to the smaller and higher values amongst *A* and *B*, respectively.

Solutions *C* and *D* are assigned the remaining ranks successively.

Ranking of predicted domain boundaries for length groups of 251–300 and 301–400 residues

Additional features are required for ranking predicted solutions for the longer proteins. For this purpose, DPCP_Phase1 was repeated on the training samples of proteins in the length group 251–300. Multiple sequence alignment of the PSIPRIED-predicted secondary structure of all the segments (of ±15 residues) around the solution "*A*" was then done using ClustalW (http://www.ebi.ac.uk/clustalw). The signature profile thus obtained was stored as *signature*_A. Similarly, multiple alignments of the segments around solutions *B*, *C*, and *D* were performed correspondingly to obtain *signature*_B, *signature*_C, and *signature*_D. The same experiment was repeated on protein sequences in the length group 301–400. A sample output of *signature*_B for a subset of the training sample in the length group 251–300 is shown in Fig. 1.

```
1gnw  ---CCCCCHHHHHHHHHHHHHHHHHHHHHHHHH----
1aw9  ---CCCCCHHHHHHHHHHHHHHHCCCHHHHHHHH----
1axd  ---CCCHHHHHHHHHHHHHHHHHHHHHHHHHHH----
1grj  -------CHHHHHHHHHHHHHHHHHHHHHHHHHCCEEC
1dow  -------HHHHHHHHHHHHHHHHHHHHHHHHHCCCHH
2trt  --CCHHHHHHHHHHHHHHHHHHHHCCCCCCCHH-----
1d1d  ---CCCCCHHHHHHHHHHHHHHCCCCCCCCCCC----
1qrj  HHCCCCCCHHHHHHHHHHHHHHHCCCCCCC-------
1dto  ----HHHHHHHHHHHHHHHHHCCCCCCCCCCCC---
```

Fig. 1 Sample output for a signature profile of multiple alignment. ClustalW output of multiple sequence alignment of segments around solution *B*. Illustration for a subset of *training sample* in the length group 251–300. "CCCCHHHHHHHHHHHHHHHHHHHHHHHCCCCC" is taken here as the signature (consensus aligned structure). Aligned end-parts are marked for a clear view

Ranking of predicted solutions

For the length group 251–300

For a given protein chain, the solutions *A* to *D* and *I* were obtained as described in DPCP_Phase1. The PSIPRED-predicted secondary structures of the segments (of ±15 residues) around the solutions *A* to *D* were aligned with the corresponding signature profiles using ClustalW. The structure with the maximum alignment score with its signature profile was marked as "*max*". If the corresponding solution was different from "*I*," then the top two ranks were assigned to these solutions (named *max* and *I*); the larger solution (position of the predicted dbp) was given the first rank and the other solution the second rank. This order of ranking was reversed if the protein length lay between 295 and 300 and the difference between the two solutions was <60. If the value of the solution *max* was the same as that of *I*, then the solution with the second highest alignment score was termed "*max*" and the top two ranks were assigned as above between this *max* and *I*.

The ranks 3 to 5 were assigned to the remaining three solutions out of *A*, *B*, *C*, and *D* in their alphabetic order.

For the length group 301–400

A similar procedure was adopted for the length group 301–400 to rank the solutions obtained by DPCP_Phase1. However, because of the greater heterogeneity of the profile-scores vs lengths, the heuristics were developed for smaller subgroups of chain-length and score-thresholds. A prominent heuristic in this case is the "midpoint_rule," instead of the rule based on "*max*" as above.

Midpoint_rule Out of the four solutions, *A*, *B*, *C* and *D*, choose the top rank as the one that is closest to the midpoint of the protein chain.

The heuristics based on profile alignment scores were then applied separately with different thresholds to the length subgroups [301 to 320] to [381 to 400] for ranking the remaining solutions by the "*max*" criteria defined above.

Discontinuous domains As 20 % of the protein sequence at each end (the N and C termini) is masked in the DPCP algorithm, we consider the following additional heuristics for proteins with two discontinuous domains.

Masked-portion heuristics

1. For the length group 80–250, if no DGS solution is found in the masked portion, or if, in the case of the N terminus, this solution is ≥40, predict the midpoint of the masked portion as a dbp.
2. For the length group 251–300, if no DGS solution is found in the masked portion at the N terminus, then predict the sequence position at one-third of this portion as a dbp. If a DGS solution lies in the masked portion at the C terminus, predict this as a dbp; if there is no DGS solution in this portion, then predict the midpoint of this portion as a dbp.
3. For the length group 301–400, if no DGS solution is found in the masked portion at the N terminus, then predict the sequence position at one-third of this portion as a dbp. If there is no DGS solution in the masked portion at the C terminus, then predict a dbp as being equal to $m-\gamma(L)$, where m denotes the midpoint of this masked segment, L is the protein length, and $\gamma(L)$ is a heuristic parameter estimated from the training sample as a multistep hop function taking uniform random values of [3.5, 8.5].

Computational experiments on different statistical designs of *training samples* showed an interesting behavior of the other dbp of discontinuous proteins. The following ranking rule was found to be significantly better than ranking by profile scores (cf. "Ranking of predicted domain boundaries for length groups of 251–300 and 301–400 residues"): rank the solutions obtained at DPCP_Phase1 in order from top to fifth as *I*, *A*, *D*, *C*, *B* for the proteins in the length group 80–250; *I*, *A*, *C*, *B*, *D* for the proteins in the length group 250–300; and *I* followed by ranking according to closeness to the midpoint for the proteins in the length group 301–400.

Results

As our method (DPCP) has direct bearing upon DGS, we have compared its performance with that of the latter for the same set of validation samples. DomSSEA also uses secondary-structure information and has, until now, given the best improvements over DGS. We have also tested the performance of DomSSEA on the same samples.

Validation results for two-continuous-domain class

The accuracy of prediction by different combinations of labeled solutions shows that solutions labeled *A* and *B* together performed better in validation samples than those labeled *C*, *D*, etc. The accuracies, at resolution levels ±10, ±15, and ±20, of predictions using these solutions, together with the solution labeled *I*, are found to be 73.5, 84.7, and 95.3 %, respectively, for length group G1 (length 80–250). These measures for G2 (length 251–300) are 47.4, 52.6, and 73.7 %, respectively, and for G3 (length 301–400), are 38.5, 53.8, and 62.8 %, respectively. With the incorporation of the solutions labeled *C* and *D*, the overall accuracy percentages improve to 79.1, 88.4, and 97.7 for G1; 66.7, 72.0, and 87.7 for G2; and 50.0, 68.0, and 75.6 for G3. However, individual labels do not show any significant performance; moreover, no consistency is found in ordering their performance according to their labels. Hence, a separate ranking procedure is sought in DPCP_Phase2.

Testing of solutions ranked by the algorithms described in the section "DPCP_Phase2" shows that the top1 and the

overall predictions of DPCP are significantly better than those of DGS and DomSSEA in most cases, and are comparable in the case of the other predictions. The percentage accuracies of the three methods for proteins in the same validation sample are shown in Table 2.

Percentage accuracy plots, as continuous functions of protein chain length, are shown in Fig. 2a,b. It can easily be seen from each pair of curves for top1, top2, and top5 that the performance of DPCP is better than that of DGS and DomSSEA. DPCP is also more consistent (fewer fluctuations) and has much lower variance of percentage accuracy from protein to protein.

Validation results for two-discontinuous-domain class

This improvement in the overall, as well as the top-ranked, predictions by DPCP as compared to DGS is remarkable in the case of discontinuous-domain proteins. Here, the resolution level ±21 is used for accurate prediction, as below this, the performance of DGS was still worse.

For the prediction of at least one dbp, the percentage accuracies of the *top_min* (i.e., the minimum number of solutions required for prediction of all dbps) and overall solutions by DGS are 88.3 and 100 for G1, 62.5 and 79.2 for G2, and 56.7 and 68.3 for G3, respectively. Those of DomSSEA are 32.9 and 44.3 for G1, 15.1 and 30.3 for G2, and 26.7 and 45.1 for G3. The accuracy percentages of the *top_min* and overall solutions by our method DPCP are 90.0 and 91.7 for G1, 100.0 and 100.0 for G2, and 95.0 and 95.0 for G3.

For the prediction of all the dbps, the accuracy percentages of the *top_min* and overall solutions are found to be the following: DGS prediction accuracies (%) are 16.7 and 36.7 for G1, 8.3 and 12.5 for G2, and 15.0 and 15.0 for G3, respectively. DomSSEA prediction accuracies (%) are 0.0 and 0.0 for G1, 3.2 and 8.1 for G2, and 7.9 and 13.2 for G3. The respective accuracies (%) of DPCP are 73.3 and 78.3 for G1, 79.2 and 91.7 for G2, and 60.0 and 71.7 for G3.

Interestingly, more than 85 % of the accurate predictions of DPCP satisfy the resolution level ±15. The percentage accuracy curves as functions of protein lengths are shown in Fig. 3a,b. The graphs clearly show a significantly better and steady (consistent) performance of DPCP against DGS and DomSSEA. The degree of variation in this reliability measure from protein to protein is also lower compared to the latter methods.

Overall superiority

On an average, the accuracy of total (top five) predictions by DPCP improved the average accuracy of the top5 solutions of DGS from 71.74 to 82.88 % in the case of two-continuous-domain proteins, and from 21.38 to 80.56 % for two-discontinuous-domain proteins. Considering only the topmost solution, the gains in accuracy are from 0 to 72.74 % for two-continuous-domain proteins with chain lengths of up to 300 residues, and from 0 to 62.85 % for those of up to 400 residues; for two discontinuous domains, the topmost (*top_min*) solutions of DPCP improve the average accuracy of DGS prediction from 12.5 to 76.3 % in proteins with chain-lengths of up to 300 residues, and from 13.33 to 70.84 % for longer proteins.

In our computational experiments with DomSSEA on the proteins of the same validation samples, the average accuracies of the topmost solutions were found to be 61 and 52 % for samples of proteins of up to 300 and 400 residues, respectively, in the case of continuous domains; those for discontinuous domains are 28 and 21 %, respectively.

The average prediction accuracy of the program PPRODO [8] is reported to be around 66 % for two-continuous-domain proteins. This value is comparable with that of DPCP (67.8 %). However, PPRODO is not applicable to the discontinuous-domains case. Moreover, because it uses cascades of neural networks, PPRODO's computational complexity is significantly greater than that of DGS, DPCP, and DomSSEA.

Discussion

DPCP is computationally simple, like DGS, and provides significant refinement in the performance of the latter by making use of some conserved local structural folds. At the most, five predictions are made for the likely dbps. The

Table 2 Percent accuracy of prediction of dbps in *2d* proteins by our method (DPCP), DGS, and DomSSEA

Protein length group	Method	% Accuracy of prediction (at resolution level ±20) of the top ranked solutions		
		Top1	Top2	Top5
80–250 (G1)	DGS	0.0	62.3	98.6
	DomSSEA	61.6	64.9	70.8
	DPCP	82.3	90.7	97.7
251–300 (G2)	DGS	0.0	36.8	80.7
	DomSSEA	35.7	41.1	42.8
	DPCP	63.2	73.7	87.7
301–400 (G3)	DGS	0.0	30.4	54.1
	DomSSEA	55.1	62.8	70.5
	DPCP	54.1	56.2	73.1

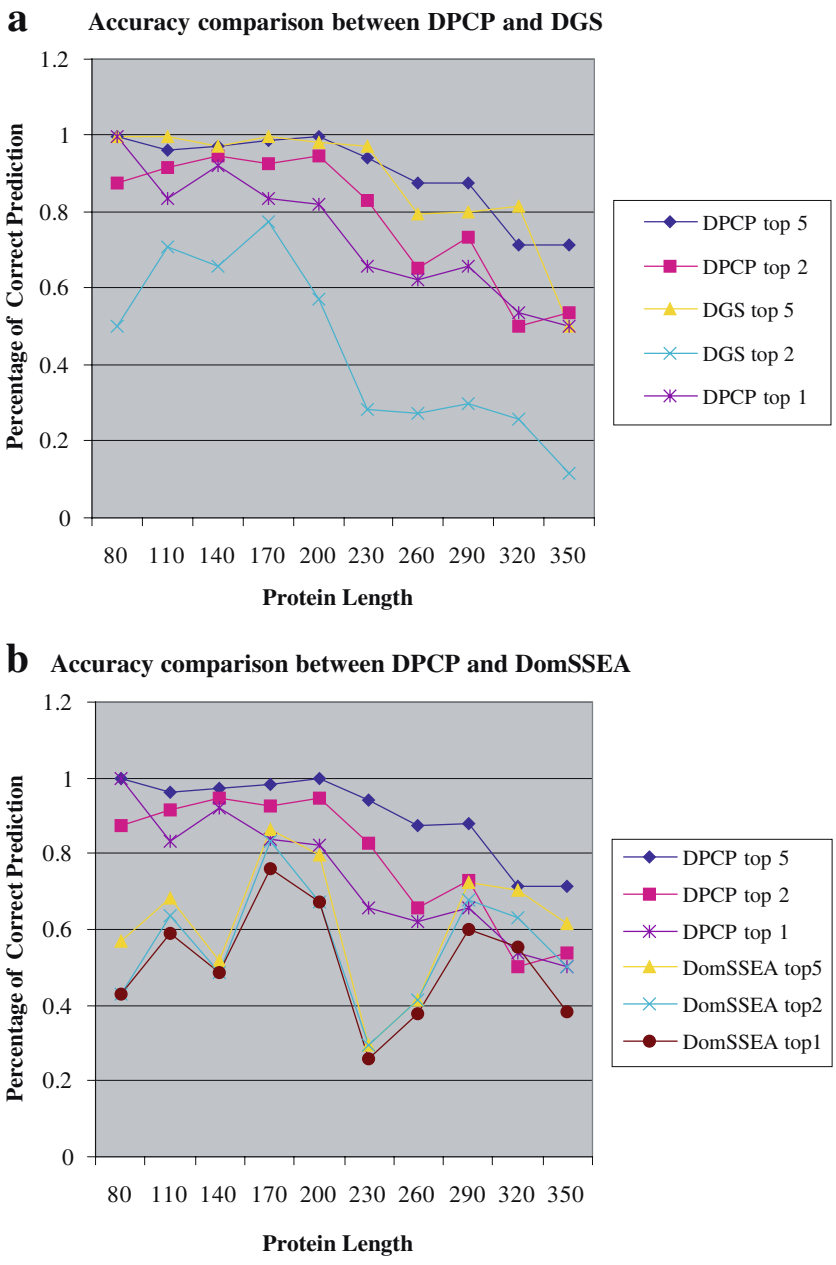

Fig. 2 Accuracy curves of DPCP for two-continuous-domain proteins. The y-axis shows % accuracy scaled on [0,1]. **a** Comparison with DGS. **b** Comparison with DomSSEA

overall performance and the accuracy of top-two- and top-one-ranked solutions, obtained by DPCP, are mostly found to be significantly better than those of the best-known existing methods/programs like DomSSEA. The performance of DPCP is also more consistent with respect to the variation in the size (length) of the protein chains. The reliability of DPCP is remarkably superior to that of the other methods when applied to proteins with discontinuous domains. DPCP is also an ab initio method, except that, at present, the secondary structures are predicted here using the PSIPRED software [11], which requires sequence homology. Nevertheless, any other comparable method of secondary-structure prediction can also be used.

In a recent study, we made use of the local folds (secondary structures) predicted by our ab initio method of protein-structure computation by nonparametric statistics and AI [12] in the prediction of a multidomain EF-hand calcium binding protein [13]. Interesting properties of the flexibility of the linker region were shown there vis-a-vis the structure (pdb_id: 1jfk) derived from NMR and in terms of some functions of this 134-residue protein of *Entamoeba histolytica*, which plays a major role in the pathogenesis of amoebiasis [14]. This prompted us to investigate further properties of domain-regions and employ them in a computationally simple algorithm that would extend the scope of protein three-dimensional-structure and function-predicting methods.

We have performed extensive data mining to explore the role of amino acid similarity clusters that incorporate important biophysical and chemical properties [15] of these building blocks of proteins; extended experiments were

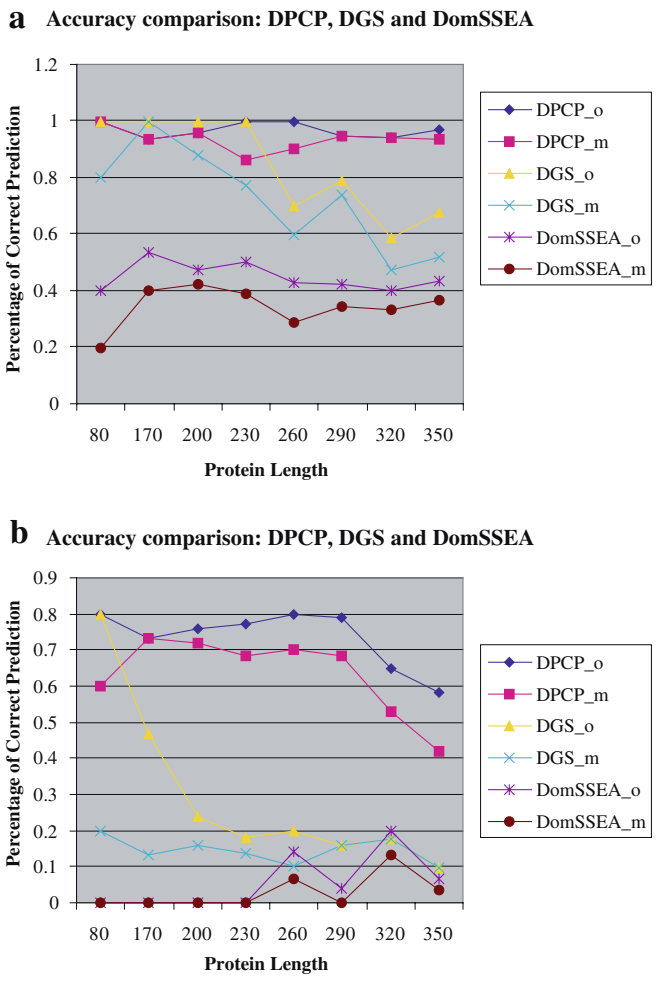

Fig. 3 Accuracy (% scaled on [0,1]) of DPCP, DGS, and DomSSEA for two-discontinuous-domain proteins. The notation *_o* denotes the consideration of overall solutions, and *_m* indicates the minimum number of top-ranked solutions required for the prediction of all dbps by the corresponding method. **a** Correct prediction for at least one dbp. **b** Correct prediction of all the dbps

also conducted using the propensities of the amino acids, computed in terms of relative likelihood, for some short motifs that were reported in early studies [16] to be common occurrences in the tertiary folds near the termini of structural domains. However, none of these succeeded in good predictions of dbps. Hence, a data-driven search for the conserved patterns of small fragments that preserve the sequential context together with the individual properties of the amino acids was tried in conjunction with the remarkable features of DGS.

Comparison of the methodologies of DPCP and DGS gives a straight implication: nature prefers not only a narrow distribution of domain size, but also certain context dependence or modularity with respect to the local folds.

Nonparametric statistical analysis of the distribution of DGS ranks or likelihood scores within and between different labels (*A*, *B*, *C*, and *D*) assigned by DPCP, in terms of patches of conserved patterns, would throw further light on the competitive or associative stochastic inter-relationship of size and secondary structure in influencing the formation of domains. Elucidation of the role of genetic recombination and of the genotype–phenotype relations in terms of the known functions of the linker and domain regions would be a possible extension of this study.

Role of heuristics

Our approach in formulating the ranking heuristics was totally data-driven and similar to the greedy-search algorithm in AI, as we wanted to identify and analyze the important factors and parameters at the primary sequence (without homology) and secondary structural levels independent of any prior assumption or constraint, to allow full flexibility and exhaustive ab initio data mining. Another reason for adopting this approach was that none of the leading predictive methods have so far been reported to be good in the case of two discontinuous domains and proteins with more than two domains.

Supervised learning algorithms in general, and heuristics derived from training samples in particular, are often thought to be biased due to "overtraining." However, these are the only possible "estimates" to extract information from the highly random and unstructured data where conventional statistical modeling fails. The labeling and ranking heuristics used here are exhaustive as they incorporate the impact of conserved patterns and their random variation with respect to the length and domain size distributions of the proteins. The statistical approach is nonparametric and jackknife-type (to deal with small samples) and the heuristics are totally data-driven, so no assumptions or influence of any specific property of the training sample subgroups would dominate. Hence, the possibility of overtraining is minimal. The consistency of our results in validation runs supports this consideration. The performance in comparison with DGS and DomSSEA further strengthens it. We will extend this study to investigate more rigorous decision criteria along the lines of Bayesian learning. The present work shows the promising potential of DPCP, and the heuristics used here indicate some interesting features, as outlined below.

The midpoint_rule together with the signature-profile-matching scores are found to be prominent in the case of ***2d*** proteins in the longer-length groups (>300 residues). However, the signature-profile-matching scores are found to be insignificant in the case of the proteins in this length group, which are ***2dd***. The masked portion heuristics and the midpoint_rule are found to be significant in efficient predictions of dbps in all the groups of proteins with discontinuous domains.

The influence of the midpoint_rule is in agreement with the uniform distribution approach of CATH, and those of DGS, DomSSEA, and the other methods compared in the performance evaluation of the latter [7]. Supporting results on protein coevolution with chaperones and protein structural duplication at the genetic control level are also cited by Wheelan et al. [2] and Marsden et al. [7]. Another notable possibility is that, as for the events of folding of globular proteins, the influence of hydrophobic cores in the

interior of the protein separated by nonhydrophobic patches on one or both sides give rise to one or more compact folds as independent globular structures or structural domains. The consistency of the radius of gyration of the globular folds as a function of chain size [17] supports this possibility. The geometrical invariances of certain functional sites [9] also show interesting possibilities for extensions in this regard. This would also explain how and why discontinuous domains are formed in chains of almost the same size as those in the continuous-domain class, in spite of the narrow domain-size distribution and the genetic events of protein duplication.

Noting that our data sets have nonredundant protein chains in each length group, and that the chains had no homology, the role of alignment scores with signature profiles of the secondary structures corresponds to a nonparametric statistical designing, which is more important in the class of long proteins because of greater heterogeneity in domain point distributions with respect to chain size and structural properties. However, the negligible influence of alignment scores and, on the contrary, the influence of masked portion heuristics in proteins with discontinuous domains, pose an intriguing query: is it the distinct distribution of hydrophobic patches along the chain that makes a difference with or without the role of the promoter gene, or is it some recombinant effect at the gene level? Nevertheless, the latter alone does not seem to be the case, as the size distributions of the segments in the discontinuous domains are found to be heterogeneous in our data with 90 % statistical confidence, which contradicts the implications discussed in the related studies so far (e.g., Wheelan et al. [2] and Marsden et al. [7]).

Large-sample data mining by CART (classification and regression trees) and the use of probabilistic relational modeling with respect to hydrophobic patch and domain-size distributions and specific genetic configurations of promotor and/or exon sequences associated with protein transcription would elucidate some of these facts and further our understanding of protein structural genomics and folding.

Acknowledgements This research is part of a software-development project sponsored by the Department of Biotechnology, Government of India. The authors thank the Department of Biotechnology for the financial support.

Appendix

1. Major steps in DPCP_0 (cf. section "DPCP_0: predictive classification of single- and two-domain proteins")

An *input set*, a random subset of the *training sample* (in the desired protein length group of 70–250, 251–300, or 301–400), is chosen so that each domain-class of interest contains comparable representations.

The following steps are executed for each of the above-mentioned length groups of proteins:

The prior probability of each domain-class (single—***1d***, two continuous domains—***2d***, and two discontinuous domains—***2dd***) is computed as its relative frequency in the *input set*.

Using this *input set*, the joint and conditional probabilities of occurrence of standard tertiary motifs and the secondary structural patterns (defined in the section "DPCP_0: predictive classification of single- and two-domain proteins") in a sliding window (of size 40) portion of the protein sequence are computed.

Optimal segments and events

The segments (along protein chain) and events for which the difference in the estimated probabilities is most distinctive in the three classes are identified.

The following *events* were found to be significant in our computational experiments (Joshi and Samant, personal communication):

$$E_1\text{: No. of HLX} = 1 \; E_3\text{: No .of STR} = 1 \; E_5\text{: No .of LOOP} = 1$$
$$E_2\text{: No. of HLX} > 1 \; E_4\text{: No. of STR} > 1 \; E_6\text{: No. of LOOP} > 1$$

where HLX, STR and LOOP, which are defined below, are such that the joint probabilities of co-occurrence of these secondary motifs and the corresponding standard three-dimensional motifs (*helix*, *strand*, and *loop*, respectively) were maximum in the optimal segments near the domain regions in the *input set*.

Definition of secondary motifs

HLX represents a consecutive patch of six or more Hs predicted by PSIPRED with confidence level ≥8, STR represents a consecutive patch of two or more Es with confidence level ≥5, and LOOP represents consecutive Cs with confidence level ≥6.

The Bayes' decision function (for likelihood of domain class k, given the presence of E_i in the optimal segment of the protein chain) is computed as:

$$\psi * (k/D_i) = P(k/E_i)\, \lambda\, (E_i,\, D_i/k)$$

where k denotes the type of the domain class, viz., ***1d***, ***2d***, and ***2dd***. D_i, where i=1, ..., 6, denotes the analogs of events E_1 to E_6 for the corresponding three-dimensional motifs.

The *posterior probability*$P(k|E_i)$ is estimated from the prior and conditional probabilities and $\lambda()$ is estimated in terms of the joint probabilities of the corresponding secondary and tertiary motifs in the specified region (the probabilities functions are as estimated for the *input set*).

Classification criteria using the decision function are derived by applying it to the entire *trainingsample*. Namely, the upper and lower bounds like α_k, β_k, ξ_k, etc., are estimated.

For a new protein, the primary chain and PSIPRED-predicted secondary structure are read.

The decision function $\psi^*(k|D_i)$ is computed if E_i is present in the optimal segment of the protein chain.

Class k is predicted if this computed value lies in $[\alpha_k, \beta_k]$; in case of a tie between two classes, the odds of their relative chances are computed as the ratios of the corresponding ξ. The class the having higher odd ratio in its favor is predicted. If the odd ratio is not significant, the class with the higher *posterior probability* is chosen.

If no E_i is present in the optimal segment, the class with the higher prior probability is predicted.

2. Formulation of heuristics for DPCP_Phase 2 (cf. section "DPCP_Phase2: ranking of domain boundary predictions labeled A, B, C, D, and I")

We consider only those proteins in the *training sample* that belong to the class ***2d*** or ***2dd***.

For each protein, we take observations on the random variable X_0, where X_0=the minimum distance of the true dbp from a predicted solution (for each protein, there are five observations, corresponding to the five predictions *A* to *D* and *I*; X_0 would correspond to the solution closest to the true dbp). Observations on the maximum distance X^* and the intermediate distances are also collected.

The following method is then applied separately—to a priori ensure the minimal heterogeneity in data—to classes ***2d*** and ***2dd*** and for the three length groups in each.

The best statistical design for analysis of the group under consideration is obtained by a *nonparametric clustering* technique to observations on (X_0, X^*). For example, the best design for the length group 80–250 in class ***2d*** thus obtained is two clusters of proteins having lengths 80–240 and 241–250. For each such *cluster*, heuristics are developed to rank the five predicted solutions according their relative likelihood of being the solutions corresponding to X_0 and to the successive higher values until X^*.

As no optimal *clustering* is obtained for longer proteins using observations on (X_0, X^*), for these length groups, X_0 and an additional feature are also used viz., the alignment profile score (described in "Ranking of predicted domain boundaries for length groups of 251–300 and 301–400 residues") for *clustering*. The length intervals of the subgroups thus obtained are 5 to 20; e.g., 295–300, 301–320, etc.

Noting that the proteins in the ***2dd*** class would have at least two dbps, other heuristics (e.g., the masked-portion heuristics) are also devised using the relative frequency of occurrence of a dbp near the two ends of the protein chain.

References

1. Xu D, Xu Y, Uberbacher EC (2000) Curr Prot Peptide Sci 1: 1–21 (http://www.bentham.org/cpps/ContentAbstract.htm)
2. Wheelan SJ, Marchler-Baucer A, Bryant SH (2000) Bioinformatics 16:613–618
3. Suyama M, Ohara O (2003) Bioinformatics 19:673–674
4. Tanaka T, Kuroda Y, Yokoyama S (2003) J Struct Funct Genomics 4:79–85
5. Sowdhamini R, Rufino SD, Blundell TL (1996) Fold Des 1:209–220
6. Jones S, Stewart M, Michie A, Swindells MB, Orengo C, Thornton JM (1998) Protein Sci 7:233–242
7. Marsden RL, McGuffin LJ, Jones DT (2002) Protein Sci 11:2814–2824
8. Sim J, Kim SY, Lee J (2005) Proteins: Struct Funct Genet 59:627–632
9. Tendulkar AV, Wangikar PP, Sohoni MA, Samant VV, Mone CY (2003) J Mol Biol 334:157–172
10. Tendulkar AV, Joshi AA, Wangikar PP, Sohoni MA (2004) J Mol Biol 338:611–629
11. McGuffin LJ, Jones DT (2000) Bioinformatics 16:404–405
12. Joshi RR, Jyothi S (2003) Comput Biol Chem 27:241–252
13. Jyothi S, Mustafi SM, Chary KVR, Joshi RR (2005) J Mol Mod 11:481–488
14. Joshi RR, Jyothi S (2005) IBS conference & national symposium on recent trends in molecular and medical biophysics. IBS, Pune, pp 22–25
15. Stanfel LE (1996) J Theoret Biol 183:195–205
16. Hutchinson EG, Thornton JM (1996) Protein Sci 5:212–220
17. Batencourt MR, Skolnick J (2001) Biopolymers 59:305–309