Gene Ontology Based Function Prediction of Human Protein Using Protein Sequence and Neighborhood Property of PPI Network

Sovan Saha, Piyali Chatterjee, Subhadip Basu and Mita Nasipuri

Abstract Predicting functions of protein from its amino acid sequence and interacting protein partner is one of the major challenges in post genomic era compared with costly, time consuming biological wet lab techniques. In drug discovery, target protein identification is important step as its inhibition may disturb the activities of pathogen. So, the knowledge of protein function is necessary to inspect the cause of diseases. In this work, we have proposed two function prediction methods FunPred1.1 and FunPred1.2 which use neighbourhood analysis of unknown protein empowered with Amino Acid physico-chemical properties. The basic objective and working of these two methods are almost similar but FunPred1.1 works on the entire neighbourhood graph of unknown protein whereas FunPred1.2 does same with greater efficiency on the densely connected neighbourhood graph considering edge clustering coefficient. In terms of time and performance, FunPred1.2 achieves better than FunPred1.1. All the relevant data, source code and detailed performance on test data are available for download at FunPred-1.

S. Saha (✉)
Department of Computer Science and Engineering, Dr. Sudhir Chandra
Sur Degree Engineering College, Dumdum 700 074, Kolkata, India
e-mail: Sovan.Saha12@gmail.com; sovansaha12@gmail.com

P. Chatterjee
Department of Computer Science and Engineering, Netaji Subhash
Engineering College, Garia 700152, Kolkata, India
e-mail: piyali.gini@gmail.com; chatterjee_piyali@yahoo.com

S. Basu · M. Nasipuri
Department of Computer Science and Engineering, Jadavpur University,
Kolkata 700032, India
e-mail: Subhadip.Basu@cse.jdvu.ac.in; subhadip@cse.jdvu.ac.in

M. Nasipuri
e-mail: Mita.Nasipuri@cse.jdvu.ac.in; mitanasipuri@yahoo.com

S.C. Satapathy et al. (eds.), *Proceedings of the 5th International Conference on Frontiers in Intelligent Computing: Theory and Applications*, Advances in Intelligent Systems and Computing 516, DOI 10.1007/978-981-10-3156-4_11

Keywords Protein function prediction • Neighbourhood analysis • Protein protein interaction (PPI) network • Functional groups • Relative functional similarity • Edge clustering coefficient • Cellular component • Biological process • Molecular function • Physico-chemical properties • Gene Ontology(GO)

1 Introduction

Proteins executes vital functions in essentially all biological processes. Computational methods like gene neighborhood, sequence and structure, protein-protein interactions (PPI) etc. have naturally created a larger impact in the field of protein function prediction than the biological based experimental methods. Unknown protein function predicted from protein interaction information is an emerging area of research in the field of bioinformatics. In this approach functions of unannotated proteins are determined by utilizing their neighborhood properties in PPI network on the basis of the fact that neighbors of a particular protein have similar function.

In the work of Schwikowski [1] at first most frequent occurrence of k functional labels are identified. Then a simple counting technique is used to assign k functions to the unannotated protein based on the identification. Though the entire methodology is not too much complex in execution but the fact that the entire network has not been considered cannot be denied. Besides confidence score also play an important role in predicting functional annotations which is also missing in this work. This deficiency of assignment of confidence score has been erased in the work of Hishigaki et al. [2]. Here annotations of k functions to the unannotated protein P is dependent on k largest $chi-square$ scores which is defined as $\frac{(n_f - e_f)^2}{e_f}$, where n_f is the count of proteins belonging to the n-neighborhood of the protein P that have the function f and e_f is the expectation of this number based on the number of occurrences of$f$among all proteins available in the entire network. While on the other hand, the exploitation of the neighborhood property of PPI network up to the higher levels has been executed in the work of Chen et al. [3]. Whereas Vazquez et al. [4] annotate a protein to a function in such a way that the connectivity of the allocated protein to that function is maximum. An identical technique on a collection of PPI data as well as on gene expression data is applied by Karaoz et al. [5]. Nabieva et al. [6] applies a flow based approach considering the local as well as global properties of the graph. This approach predicts protein function based on the amount of flow it receives during simulation. It should be noted here that each annotated protein acts as the source of functional flow. While the theory of Markov random field has been reflected in the work of Deng et al. [7] where the posterior probability of a protein of interest is estimated. Letvsky and Kasif [8] use totally a different approach by the application of binomial model in unknown protein function prediction. Similarly, Wu et al. [9] includes the summation of both protein structure and probabilistic approach in this field of study. In the work of Samanta et al. [10], a network based statistical algorithm is proposed, which assumes that if two proteins share significantly larger number of common interacting partners they share a common functionality. Arnau et al. [11] proposed another application named as UVCLUSTER which is based on bi-clustering. This application iteratively explored distance datasets. In the early stage, Molecular Complex Detection (MCODE) is executed by Bader and Hogue [12] where identification of dense regions takes place according to some heuristic parameters. Altaf-ul-Amin et al. [13] also use a clustering approach. It starts from a single node in a graph and clusters are gradually grown until the similarity of every added node within a cluster and density of clusters reaches a certain limit. Graph clustering approach is used by Spirin and Mirny [14] where they detect densely connected modules within themselves as well as sparsely connected with the rest of the network based on super paramagnetic clustering and Monte Carlo algorithm. Theoretical graph based approach is observed in the work of Pruzli et al. [15] where clusters are identified using Leda's routine components and those clusters are analyzed by Highly Connected Sub-graphs (HCS) algorithm. While the application of Restricted Neighborhood Search Clustering algorithm (RNCS) is highlighted in the work of King et al. [16]. The interaction networks are partitioned into clusters by this algorithm using a cost weightage mechanism. Filtering of clusters is then carried out based on their properties like size, density etc.

This survey highlights the fact that there is an opportunity for inclusion of domain as well as some other related specific knowledge like protein sequences to enhance the performance of protein function prediction from protein interaction network. Motivated by this fact, a neighborhood based method has been proposed for predicting function of an unannotated protein by computing the neighborhood scores on the basis of protein functions and physico-chemical properties of amino acid sequences of proteins. The unannotated protein is associated with the function corresponding to highest neighborhood score.

1.1 Dataset

We have used the Gene Ontology (GO) dataset of human obtained from UniProt. The dataset is available at FunPred-1. Three categories: Cellular-component, Molecular-function and Biological-process are involved in the GO system. In this system, each protein may be annotated by several GO terms (like GO: 0000016) in each category. So, here, at first we have ranked every GO terms of 3 categories based on the maximum number of occurrences in each of them. Then 10 % of proteins belonging to the top 15 GO terms in each of three categories are selected as unannotated while the remaining 90 % proteins are chosen as training samples using random sub-sampling technique. Since we have considered both *Level*-1 and *Level*-2 neighbors, the protein interaction network formed for each protein in any functional group is large and complex. Therefore, in the current experiment we consider only 10 % of available proteins in each functional group as test set. Table 1 show the detailed statistics of the train-test dataset for the three GO categories. While overall

Table 1 Distribution of proteins and protein pairs in 3 functional categories in GO based Human dataset, considered under the current experiment

Organism	Number of proteins	No. of interactions	GO terms	Cellular component	Molecular function	Biological process
Human	2577	3329	3730	522	717	2491
	Cellular component		*Molecular function*		*Biological process*	
	Selected unannotated proteins	Annotated proteins	Selected unannotated proteins	Annotated proteins	Selected unannotated proteins	Annotated proteins
	846	1731	765	1812	1216	1361
	Total selected unannotated proteins in entire GO : 846 + 765 + 1261 = 2872					
	Total annotated proteins in entire GO: 1731 + 1812 + 1361 = 4904					

protein interaction network of the three functional categories along with known (marked blue) and unannotated proteins (marked yellow) with their respective result comparison by FunPred 1 has been highlighted here.

2 Related Terminologies

In both FunPred 1.1 and FunPred 1.2, we have used four scoring techniques: Protein Neighborhood Ratio Score $\left(\text{Pscore}^{1(=1,2)}\right)$[17], Relative functional similarity $\left(W_{u,v}^{l(=1,2)}\right)$ [17, 18], Proteins path-connectivity score $\left(Q_{u,v}^{l(=1,2)}\right)$ [17, 19] and physico-chemical properties score $\left(\text{PCP}_{\text{score}}^{l(=1,2)}\right)$ [20]. $\text{PCP}_{\text{score}}^{l(=1,2)}$ is incorporated since sequences of amino acid of each protein also plays a vital role in unknown protein function prediction. While in FunPred 1.2, we have used one additional feature Edge Clustering Coefficient $\left(\text{ECC}_{u,v}^{l(=1,2)}\right)$ [21] to find densely populated region in the network. All the other relevant graphical terms and properties are described in our previous work [17, 22].

3 Proposed Method

Two methods [17] have been proposed for unannotated protein function prediction. Uniqueness can be defined in the aspect that the selection of the neighborhood of the unannotated proteins in both these two methods differs over the different aspects of neighborhood properties defined in the previous section. The first method FunPred 1 is described below:

3.1 FunPred 1.1

FunPred 1.1 [17] uses the combined score of neighborhood ratio, proteins path connectivity, physico-chemical property score and relative functional similarity. Now, this method always focuses in identifying the maximum of the summation of four scores thus obtained in each level and assign the unannotated protein to the corresponding functional group (GO term) of the protein having the maximum value. Given G_P', a sub graph consisting of any proteins (nodes) of set $FC = \{FC_1, FC_2, FC_3\}$; where, FC_i represents a particular functional category, this method annotates proteins belonging to the set of un-annotated proteins P_{UP} to any GO term of set FC. Steps of FunPred 1.1 are described as Algorithm 1.

Algorithm 1 Basic methodology of FunPred 1.1

```
Input:Unannotated protein set $P_{UP}$.
Output:The proteins of the set $P_{UP}$ gets annotated to any
functional group (GO term) ofset FC.
Step 1: Any protein from set $P_{UP}$ is selected.
Step 2: Count $Level-1$ and $Level-2$ neighbors of that
protein in $G_P'$ associated with set FC.
Step 3: Compute $P^{l(=1,2)}_{FC_{i(=1,\ldots,3)}}$ for each GO term in set FC and
assign this score to eachprotein
$\left(Pscore^{l(=1,2)}\right) \in P_A$, belonging to the respective
functionalcategory.
Step 4: Compute $Q^{l(=1,2)}_{u,v}$, $W^{l(=1,2)}_{u,v}$ for each edge in $Level-1$
and $Level-2$.
Step 5:Obtain neighborhood score i.e.
```

$$N^l_{(FC_K)} = Max\left(\left(max(Pscore^1 + Q^1_{u,v} + W^1_{u,v} + ECC^1_{u,v} + PCP^1_{score})\right),\right.$$
$$\left.\left(max(Pscore^2 + Q^2_{u,v} + W^2_{u,v} + ECC^2_{u,v} + PCP^2_{score})\right)\right)$$

```
Step 6: The unannotated protein from the set $P_{UP}$ is
assigned to the GO term belonging to $FC_K$.
```

3.2 FunPred 1.2

In FunPred 1.1, all *Level*-1 neighbors and *Level*-2 neighbors belonging to any GO term of 3 functional categories are considered for any unannotated protein. Neighborhood property based prediction is then carried out, the computation of which considers all *Level*-1 or *Level*-2 neighbors. But if the computation is confined only on significant neighbors who have maximum neighborhood impact on the target protein then exclusion of non-essential neighbors may substantially reduce the computational time which is the basis of our heuristic adopted in FunPred 1.2 [17]. So this method looks for the promising regions instead of calculating neighborhood ratios for all of them and only then the calculation of $N^{l}_{(FC_K)}$ is done. Here, at first edge clustering coefficient (ECC) of each edge in $Level-1$ and $Level-2$ (as mentioned in the earlier section) is calculated. Edges having relatively low edge clustering coefficient gets eliminated and thus the original network gets reduced upon which we will apply our previous method. Now the original FunPred-1.1 algorithm is applied on this reduced PPI network (renaming the entire modified method as FunPred 1.2). The computational steps associated with FunPred 1.2 are described as Algorithm 2.

Algorithm 2 Basic methodology of FunPred 1.2

```
Input:Unannotated protein set $P_{UP}$.
Output: The proteins of the set $P_{UP}$ gets annotated to any
functional group (GO term) of set $FC$.
Step 1:Any protein from set $P_{UP}$ is selected.
Step 2:Protein interaction network of the selected
protein has been constructed detecting its
$Level -1$ and $Level -2$ neighbors.
Step 3: Compute $ECC^{l(=1,2)}_{u,v}$ for each edge in $Level -1$ and
$Level -2$.
Step 4: Eliminate non-essential annotated proteins
(neighbors) associated with edges having lower
values of $ECC^{l(=1,2)}_{u,v}$ both in $Level -1$ and $Level -2$ thus
generating a densely connected reduced protein
interaction network.
Step 5:Count $Level -1$ and $Level -2$ neighbors of that
protein in $G'_P$ associated with set FC.
Step 6:Compute $P^{l(=1,2)}_{FC_{i(=1,\ldots,3)}}$ for each GO term in set FC and
assign this score to each protein $\left(Pscore^{l(=1,2)}\right) \in P_A$,
belonging to the respective functional group.
Step 7: Compute $Q^{l(=1,2)}_{u,v}$, $W^{l(=1,2)}_{u,v}$ for each edge in $Level -1$
and $Level -2$.
Step 8: Obtain neighborhood score i.e.

    $N^{1}_{(FC_K)} = Max(\left(max(Pscore^{1} + Q^{1}_{u,v} + W^{1}_{u,v} + ECC^{1}_{u,v} + PCP^{1}_{score})\right),$
    $\left(max(Pscore^{2} + Q^{2}_{u,v} + W^{2}_{u,v} + ECC^{2}_{u,v} + PCP^{2}_{score})\right))$

Step 9: The unannotated protein from the set $P_{UP}$ is
assigned to the GO term belonging to $FC_K$.
```

4 Results and Discussion

We have used standard performance measures, such as Precision (P), Recall (R) and F-Score (F) values for evaluating the training results for the ith functional category as described in our previous work [17]. The detailed analysis of FunPred 1.1 and FunPred 1.2 with respect to Precision, Recall and F-score values has been shown in Table 3. Functional category-wise Precision, Recall and F-scores of the two methods are given in Table 2. The average Precision of FunPred 1.2 is estimated as 0.743 (see Table 3). Although we observe relatively low values of Recall for the two methods, high Precision scores indicate that our algorithm has succeeded in generating more significant results. High F-score values have been retrieved in one functional category i.e. Molecular function. Ten percent of proteins from each of the high ranking GO terms in the three functional categories are considered as unannotated proteins using random sub-sampling in both of our methods.

The performance of FunPred 1.1 has been significantly improved in FunPred 1.2 as FunPred 1.2 reduces the neighborhood network. For example, from Table 2, it can be observed that a Precision improvement of 5.2 and 9.8 % occurs in the Cellular component and Molecular function respectively in FunPred 1.2 over FunPred 1.1. In our experiment, Biological process performs worst in comparison to the other functional category. Except this category, in almost all other cases we have either achieved good prediction performance in FunPred 1.1 or obtained significant hike in performance in FunPred 1.2 in comparison to its predecessor.

To compare the performance of the current method with the other existing neighborhood analysis methods, we have identified four relevant methods and

Table 2 Evaluated results of FunPred 1.1 and FunPred 1.2 for three functional categories of GO based human dataset

Functional categories	Methods	Precision	Recall	F-Score
Cellular component	FunPred-1.1	0.662	0.602	0.631
	FunPred-1.2	0.714	0.650	0.680
Molecular function	FunPred-1.1	0.722	0.725	0.724
	FunPred-1.2	0.820	0.823	0.821
Biological process	FunPred-1.1	0.660	0.625	0.642
	FunPred-1.2	0.695	0.657	0.676

Table 3 Recall, Precision, F-Score for FunPred 1.1 and FunPred 1.2 in accordance to Mean and standard deviation

Methods	Mean/SD	Precision	Recall	F-Score
FunPred-1.1	Mean	0.681	0.651	0.665
	Standard deviation	0.035	0.065	0.050
FunPred-1.2	Mean	0.743	0.710	0.726
	Standard deviation	0.067	0.097	0.082

compared the performances of the same on our Human dataset. More specifically we compared our work with the neighborhood counting method [1], Chi-square method [2], a recent work on Neighbor Relativity Coefficient (NRC) [19] and FS-weight based method [23].

The best performance among the four methods is the work of Moosavi et al. [19]. The NRC method generates average Precision, Recall and F-score values of 0.374, 0.434 and 0.368 respectively. The detailed result analysis of our method as highlighted in Table 3 over 15 functional groups clearly reveals the fact that our method is relatively better than the NRC based method in terms of average prediction scores. This betterment is achieved since both Level-1 and Level-2 neighbors have been considered along with the exploration of a variety of scoring techniques in the human PPI network. Not only that we have also included protein sequences, successors as well as the ancestors of a specific unannotated protein while estimating neighborhood score for unannotated protein function prediction.

The result obtained in all Chi-square methods [2] is comparatively lower than the other methods because it only concentrates only on the denser region of the interaction network. The neighborhood counting method though performs well but fails when compared to NRC, FS-weight#1 (only direct neighbors are considered) and FS-weight #1 and #2 (both direct and indirect neighbors are considered) methods since it does not consider any difference between direct and indirect neighbors. Figure 1 shows a comparative detailed analysis of the four methods (taken into consideration in our work) along with our proposed systems.

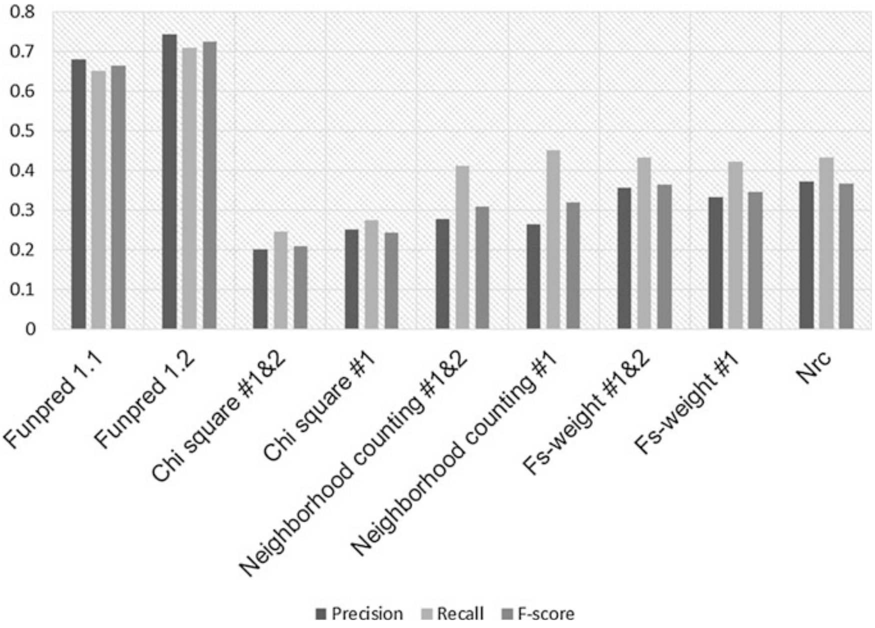

Fig. 1 Comparative analysis of other methods with our developed method FunPred 1

All these analysis show that our proposed FunPred-1 software, relatively performs much better than the other existing methods in unannotated protein function prediction. But this work is limited to only 15 high ranking GO terms/functional groups in the human PPI network, which we would like to extend for other significant GO terms as well. Simultaneously, the function prediction of our method can be well enhanced in our future work if domain-domain affinity information [24] and structure related information [25] can be incorporated.

Acknowledgments Authors are thankful to the "Center for Microprocessor Application for Training and Research" of the Computer Science Department, Jadavpur University, India, for providing infrastructure facilities during progress of the work.

References

1. B. Schwikowski, P. Uetz, S. Fields, A network of protein-protein interactions in yeast. Nat. Biotechnol. **18**, 1257–1261 (2000)
2. H. Hishigaki, K. Nakai, T. Ono, A. Tanigami, T. Takagi, Assessment of prediction accuracy of protein function from protein-protein interaction data. Yeast (Chichester, England) 18, 523–31 (2001)
3. J. Chen, W. Hsu, M.L. Lee, S.K. Ng, Labeling network motifs in protein interactomes for protein function prediction, in *IEEE 23rd International Conference on Data Engineering* (2007), pp. 546–555
4. A. Vazquez, A. Flammini, A. Maritan, A. Vespignani, Global protein function prediction from protein-protein interaction networks. Nat. Biotechnol. **21**, 697–700 (2003)
5. U. Karaoz, T.M. Murali, S. Letovsky, Y. Zheng, C. Ding, C.R. Cantor, S. Kasif, Whole-genome annotation by using evidence integration in functional-linkage networks. Proc. Natl. Acad. Sci. U. S. A. **101**, 2888–2893 (2004)
6. E. Nabieva, K. Jim, A. Agarwal, B. Chazelle, M. Singh, Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. Bioinform. **21**, i302–i310 (2005)
7. M. Deng, S. Mehta, F. Sun, T. Chen, Inferring domain–domain interactions from protein–protein interactions. Genome Res. 1540–1548 (2002)
8. S. Letovsky, S. Kasif, Predicting protein function from protein/protein interaction data: a probabilistic approach. Bioinform. **19**, i197–i204 (2003)
9. D.D. Wu, An efficient approach to detect a protein community from a seed, in *IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology* (2005), pp. 1–7
10. M.P. Samanta, S. Liang, Predicting protein functions from redundancies in large-scale protein interaction networks. Proc. Natl. Acad. Sci. U. S. A. **100**, 12579–12583 (2003)
11. V. Arnau, S. Mars, I. Marín, Iterative cluster analysis of protein interaction data. Bioinform. **21**, 364–378 (2005)
12. G.D. Bader, C.W.V. Hogue, An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinform. **27**, 1–27 (2003)
13. M. Altaf-Ul-Amin, Y. Shinbo, K. Mihara, K. Kurokawa, S. Kanaya, Development and implementation of an algorithm for detection of protein complexes in large interaction networks. BMC Bioinform. 7, doi:10.1186/1471-2105-7-207 (2006)
14. V. Spirin, L.A. Mirny, Protein complexes and functional modules in molecular networks. Proc. Natl. Acad. Sci. U. S. A. **100**, 12123–12128 (2003)

15. A.D. King, N. Przulj, I. Jurisica, Protein complex prediction via cost-based clustering. Bioinform. **20**, 3013–3020 (2004)
16. S. Asthana, O.D. King, F.D. Gibbons, F.P. Roth, Predicting protein complex membership using probabilistic network reliability. Genome Res. **14**, 1170–1175 (2004)
17. S. Saha, P. Chatterjee, S. Basu, M. Kundu, M. Nasipuri, Funpred-1: protein function prediction from a protein interaction network using neighborhood analysis cell. Mol. Biol. Lett. (2014). doi:10.2478/s11658-014-0221-5
18. X. Wu, L. Zhu, J. Guo, D.Y. Zhang, K. Lin, Prediction of yeast protein-protein interaction network: insights from the Gene Ontology and annotations. Nucleic Acids Res. **34**, 2137–2150 (2006)
19. S. Moosavi, M. Rahgozar, A. Rahimi, Protein function prediction using neighbor relativity in protein-protein interaction network. Comput. Biol. Chem. 43, doi:10.1016/j.compbiolchem.2012.12.003 (2013)
20. S. Saha, P. Chatterjee, Protein function prediction from protein interaction network using physico-chemical properties of amino acid. Int. J. Pharm. Bio. Sci. **24**, 55–65 (2014)
21. W. Peng, J. Wang, W. Wang, Q. Liu, F.X. Wu, Y. Pan, Iteration method for predicting essential proteins based on orthology and protein-protein interaction networks. BMC Syst. Biol. 6, doi:10.1186/1752-0509-6-87 (2012)
22. S. Saha, P. Chatterjee, S. Basu, M. Kundu, M. Nasipuri, Improving prediction of protein function from protein interaction network using intelligent neighborhood approach, in *International Conference on Communications, Devices and Intelligent Systems* (IEEE, 2012), pp. 604–607
23. H.N. Chua, W.K. Sung, L. Wong, Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. Bioinform. **22**, 1623–1630 (2006)
24. P. Chatterjee, S. Basu, M. Kundu, M. Nasipuri, D. Plewczynski, PSP_MCSVM: brainstorming consensus prediction of protein secondarystructures using two-stage multiclass support vector machines. J. Mol. Model. **17**, 2191–2201 (2011)
25. P. Chatterjee, S. Basu, J. Zubek, M. Kundu, M. Nasipuri, D. Plewczynski, PDP-CON: prediction of domain/linker residues in protein sequences using a consensus approach. J. Mol. Model. doi:10.1007/s00894-016-2933-0 (2016)