Amino Acids (2008) 35:599–605
DOI 10.1007/s00726-008-0085-y

ORIGINAL ARTICLE

DPROT: prediction of disordered proteins using evolutionary information

Deepti Sethi · Aarti Garg · G. P. S. Raghava

Received: 19 February 2008 / Accepted: 17 March 2008 / Published online: 19 April 2008

Abstract The association of structurally disordered proteins with a number of diseases has engendered enormous interest and therefore demands a prediction method that would facilitate their expeditious study at molecular level. The present study describes the development of a computational method for predicting disordered proteins using sequence and profile compositions as input features for the training of SVM models. First, we developed the amino acid and dipeptide compositions based SVM modules which yielded sensitivities of 75.6 and 73.2% along with Matthew's Correlation Coefficient (MCC) values of 0.75 and 0.60, respectively. In addition, the use of predicted secondary structure content (coil, sheet and helices) in the form of composition values attained a sensitivity of 76.8% and MCC value of 0.77. Finally, the training of SVM models using evolutionary information hidden in the multiple sequence alignment profile improved the prediction performance by achieving a sensitivity value of 78% and MCC of 0.78. Furthermore, when evaluated on an independent dataset of partially disordered proteins, the same SVM module provided a correct prediction rate of 86.6%. Based on the above study, a web server ("DPROT") was developed for the prediction of disordered proteins, which is available at http://www.imtech.res.in/raghava/dprot/.

Keywords Disorder proteins · Position-specific scoring matrices · Support vector machines · Amino acid composition · Web server

D. Sethi · A. Garg · G. P. S. Raghava (✉)
Scientist and Head Bioinformatics Centre,
Institute of Microbial Technology, Sector 39A,
Chandigarh, India
e-mail: raghava@imtech.res.in
URL: http://www.imtech.res.in/raghava/

Introduction

The knowledge of three-dimensional (3-D) structure of a protein is essential in order to deduce its biological function. Since, the prediction of secondary structure is an intermediate step in structure determination, our group has developed a number of secondary and supersecondary structure prediction methods previously (Kaur and Raghava 2002, 2003, 2004a, 2004b; Kumar et al. 2005).

However, the past few years have seen a growth interest in structural studies of proteins, focusing comprehensively on the study of proteins which are structurally disordered, often termed "disordered proteins." These proteins have been gaining a great deal of attention from biologists, since their involvement in various physiological disorders which could be protein deposition diseases, such as Alzheimer's and Parkinson's diseases, has become evident (Fink 2005). Recently, the implication of disordered proteins in the disease associated with single amino acid polymorphism (SAP) (Zhi-Qiang et al. 2007) has enhanced the concerns of biologists regarding these proteins. SAPs caused by the substitution of one amino acid for another can alter protein structure and function as well as its solubility and stability, and may lead to disease; see Yip et al. 2004. In their study they used the disordered regions in proteins as biologically informative attributes to find out whether the SAPs are located in structurally disordered regions. The interest in disordered proteins also stems from the important functions in which they participate. These can be broadly classified into molecular recognition, molecular assembly/disassembly, protein modification and entropic chain activities (Dunker et al. 2001, 2002).

From a structural point of view, disordered proteins, or disordered regions, are those which lack a specific tertiary structure and are composed of an ensemble of conformations, usually with distinct and dynamic Φ and Ψ (Fink 2005). These proteins in their purified state at neutral pH have either been shown experimentally or are predicted to lack an ordered structure. The existence of disorder is determined by the overall protein dynamics rather than by local secondary structure (Radivojac et al. 2004). These proteins are also referred to as "natively unfolded" (Weinreb et al. 1996) or "intrinsically unstructured" (Wright et al. 1999).

The significance of studying disordered proteins lies in the important attributes they offer, such as high average flexibility index values (Vihinen et al. 1994), low sequence complexity (Romero et al. 1999, 2001) as estimated by Shannons entropy, low aromatic content (Xie et al. 1998) and high specificity coupled to modest binding affinities.

Several predictors have already been developed for predicting disordered proteins/regions, such as PONDR (Jones and Ward 2003), DISOPRED2 (Ward et al. 2004), GlobPlot (Linding et al. 2003a), DISEMBL (Linding et al. 2003b), FoldIndex (Sussman et al. 2005) and RONN (Yang et al. 2005), etc. All of these predictors exploit various attributes of the protein sequence, such as amino acid compositions, flexibility, charge, hydropaths, PSIBLAST profiles, propensities for secondary structure and random coils, etc.

On the other hand, IUPRED (Dosztanyi et al. 2005), which is based on inter-residue interactions, predicts regions that lack a well-defined 3-D structure under native conditions, whilst FoldUnfold (Galzitskaya et al. 2006) predicts disordered regions by estimating the number of contacts of the whole protein. Recently, a predictor called POODLE has been developed (Shimizu et al. 2007), which can predict disordered proteins with a high sensitivity value of 72.3% and an accuracy of 97.7%. POODLE is based on Joachims' spectral graph transducer (SGT), which is a binary classification based on semi-supervised learning. Despite gaining such a high prediction accuracy, the method seems to be insensitive for partially disordered proteins. This insensitivity might be due to the utilization of a single protein feature, namely amino acid composition, for prediction.

The present study has been undertaken to further improve the prediction performance for the classification of ordered and disordered proteins by introducing new input features like secondary structure composition along with conventionally used protein features such as amino acid composition, dipeptide composition, and position-specific scoring matrices (PSSM) composition.

However, the best performance was observed for a PSSM-based module that captured the multiple sequence alignment information for the prediction of disordered proteins. The module, DPROT, has been implemented on a web server at http://www.imtech.res.in/raghava.

Materials and methods

Dataset

A representative dataset consisting of 608 proteins (526 ordered and 82 disordered proteins) was used in the present study. The same dataset was previously used to develop the POODLE web server (Shimizu et al. 2007). The raw dataset, retrieved from Disprot (version 3.3; Vucetic et al. 2005), was processed by following an intensive protocol. A dataset of 417 partially disordered proteins was also used for independent testing.

Support vector machines (SVMs)

In the present study, a highly successful machine learning technique termed a *support vector machine* (SVM) was used. SVMs are universal approximators based on statistical learning and optimization theory which support both regression and classification tasks and can handle multiple, continuous and categorical variables. To construct an optimal hyperplane, the SVM employs an iterative training algorithm which is used to minimize an error function. The SVM was implemented using *SVM_light* (Joachims 1999), which allow users to select various parameters and various kernel functions like radial basis function (RBF), linear and polynomial functions. It was observed that the RBF kernel performs better than the linear and polynomial kernels in the case of the amino acid composition-based SVM module. Thus, for all of the SVM modules developed in the present study, the RBF kernel was used.

Input feature vectors

The input protein features used in the present study for disorder protein prediction were amino acid composition, dipeptide composition, PSSM composition and secondary structure composition. More details about calculating amino acid and dipeptide compositions can be obtained from our previous works (Garg et al. 2005a; Lata et al. 2007).

Position-specific scoring matrices

The PSSM was generated using the PSI-BLAST (position-specific iterative BLAST) search with a cut-off *E* value of 0.001 against the large databases such as the nonredundant (NR) database available at NCBI, SWISS-PROT and PDB. In each of the three iterations, the PSSM is generated from multiple alignments of the high-scoring hits by calculating the position-specific scores for each position in the alignments. After three iterations, PSI-BLAST generates the PSSM with the highest score. The matrix contains 20 × N elements, where N is the length of the target sequence, and each element represents the frequency of occurrence of each of the 20 amino acids at a particular position in the alignment. Subsequently, the final PSSM was normalized using a sigmoid function. To make a SVM input of fixed length, we summed all of the rows in the PSSM corresponding to the same amino acid in the sequence, and then divided each element by the length of the sequence.

Secondary structure composition

In this work, for the first time an attempt was made to capture the predicted information for secondary states such as the coil, helix and beta sheet probabilities for each of the residues, in the form of composition values for 20 amino acids which eventually provided an input vector of 60 dimensions (i.e., 20 values for each secondary state). The secondary structure contents of ordered and disordered proteins were predicted using the *PSIPRED* (Jones 1999) software, which provides prediction probabilities for residues in three states—helices, beta-sheets and coils. It is a neural network-based secondary structure prediction method which uses multiple alignment information for the target sequence obtained from PSI-BLAST.

Performance evaluation

All SVM models developed in this study were evaluated using a five-fold cross-validation technique (Chou and Shen 2007b). In this technique, the dataset was partitioned randomly into five equally sized sets. The training and testing was carried out five times, each time using one distinct set for testing and the remaining four sets for training (Bhasin and Raghava 2004).

A standard set of parameters was used to evaluate the performances of the various methods developed in this study. A brief description of these parameters follows.

The *sensitivity* is the percentage of disordered proteins correctly predicted to be disordered (p), as shown in the following equation (u is the number of under-predicted sequences):

$$\text{Sensitivity} = \left(\frac{p}{p+u}\right) \times 100 \quad (1)$$

The *specificity* is the percentage of ordered proteins correctly predicted to be ordered (n), as shown in the following equation (o is the number of over-predicted sequences):

$$\text{Specificity} = \left(\frac{n}{n+o}\right) \times 100 \quad (2)$$

The *accuracy* is the percentage of correctly predicted disordered and ordered proteins among the total number of protein sequences (t):

$$\text{Accuracy} = \left(\frac{p+n}{t}\right) \times 100 \quad (3)$$

Matthew's Correlation Coefficient (MCC) of 1 corresponds to a perfect prediction, whereas 0 corresponds to a completely random prediction:

$$\text{MCC} = \frac{pn - ou}{\sqrt{(p+o)(p+u)(n+o)(n+u)}} \quad (4)$$

Results and discussion

Composition-based SVM modules

First, the amino acid compositions of all disordered as well as ordered protein sequences were computed. Figure 1 depicts the compositional differences between them. The disordered proteins were found to have low contents of aromatic amino acids. In addition, low amounts of Cys, His and high amounts of Glu, Asp, Ser and Lys were also found to be associated with disordered proteins. Further, using amino acid composition as an input feature for the training

Fig. 1 Comparison of the amino acid compositions of ordered proteins and disordered proteins

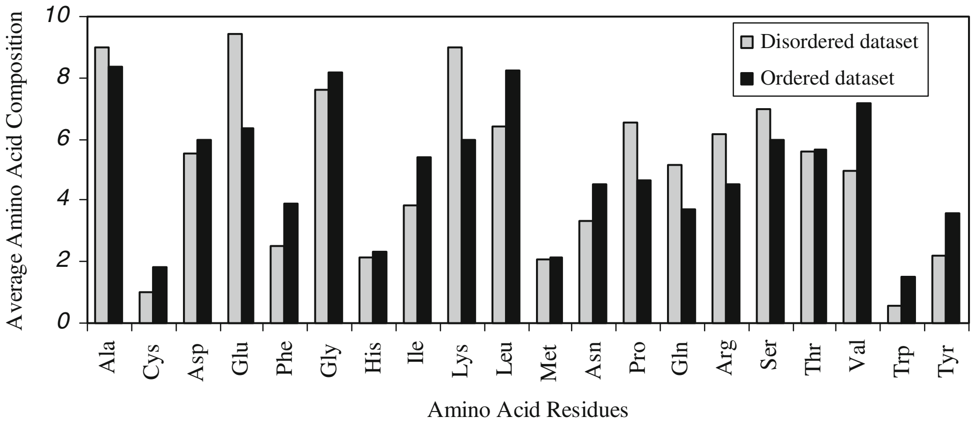

Table 1 Performances of individual SVM modules in predicting disordered proteins

Feature used	Sensitivity	Specificity	Accuracy	MCC
AA	75.6	97.1	94.2	0.75
DIPEP	73.2	92	89.5	0.60
PSSM	76.8	97.0	94.2	0.75
SS	76.8	97.7	94.9	0.77

AA, amino acid composition; *DIPEP*, dipeptide composition; *PSSM*, position-specific scoring matrices; *SS*, secondary structure composition

of SVM modules, we achieved sensitivity, specificity and an accuracy of 75.6, 97.1, and 94.2%, respectively, along with MCC value of 0.75 (RBF kernel, $g = 25$, $c = 1$, $j = 3$), as shown in Table 1.

Similarly, dipeptide composition, another protein attribute, was also used to generate the SVM-based module used to predict disordered proteins. Although dipeptide composition captures information about the local order as well as amino acid composition, a low performance using the RBF kernel ($g = 15, c = 1, j = 15$) was obtained. The SVM-based module attained sensitivity, specificity, accuracy and MCC values of 73.2, 92.0, 89.5% and 0.60, respectively (Table 1). This low performance might be due to the lack of local order in disordered proteins, and hence the information was not fully exploited when predicting the disordered proteins.

Further, attempts were then made to improve the performance using composition for a PSI-BLAST-generated PSSM against the NR database. Previously, the same input feature has been successfully applied to improve the prediction accuracy of the secondary structure, solvent accessibility and subcellular localization (Jones 1999; Rashid et al. 2007; Garg et al. 2005b; Xie et al. 2005); however, the present approach differs mainly in that an input vector of fixed length is generated for training. For instance, Pseudo-PSSM was used to predict the membrane protein types (Chou and Shen 2007a) and the enzyme main-functional and sub-functional classes (Shen and Chou 2007a) as well as the protein subnuclear localization (Shen and Chou 2007b) instead of the 400 dimensional input feature vector used in the present study. The SVM module based on PSSM composition yielded slightly better results than the amino acid-based module. Here, using the RBF kernel, we obtained sensitivity, specificity, accuracy and MCC values of 76.8, 97.0, 94.2% and 0.75, respectively.

Furthermore, a secondary structure composition-based SVM module was developed using an input feature vector of 60 dimensions. This module was found to outperform the other modules previously developed in this study. This approach with novel input features yielded sensitivity, specificity, accuracy and MCC values of 76.8, 97.7, 94.9% and 0.77 ($g = 5$, $c = 5$, $j = 3$), respectively.

Hybrid modules

Hybrid approach-based SVM modules were also developed using combinations of individual feature modules in order to get high sensitivity without losing much specificity or high specificity with a reasonable percentage coverage for the disorder protein prediction method. Each approach had its own limitations, as some provided high sensitivity values but low specificity and vice versa.

The results for the hybrid approach-based SVM modules developed in the present study are shown in Table 2. The hybrid modules were generated using combinations of two or three and even all of the individual modules (amino acid composition, dipeptide composition, PSSM composition and secondary structure composition) developed in the present study. The dihybrid module developed using amino acid and PSSM composition as input features provided sensitivity, specificity, accuracy and MCC values of 76.8, 97.0, 94.2% and 0.75, respectively, with the RBF kernel ($g = 10$, $c = 275$, $j = 3$), a similar performance to that obtained for the PSSM-based SVM module. Similarly, other SVM-based dihybrid modules encapsulating the information on amino acid composition along with the secondary structure composition, and another module developed using the secondary structure and PSSM information showed slightly improved results by attaining sensitivity, specificity, accuracy and MCC values of 78.0, 97.0, 94.4% and 0.76, respectively. Finally, combining all the individual SVM-based modules such as those for amino acid, dipeptide, PSSM and secondary structure composition resulted in an input vector with 880 dimensions. However, the training of an SVM model with a vector of 880 dimensions was observed to be computationally very expensive, and unfortunately failed to produce any improvement in the performance. Thus, further endeavors were made to develop a module which could improve on for the shortcomings of hybrid modules.

Table 2 Performances of hybrid and cascade modules in predicting disordered proteins

Feature used	Sensitivity	Specificity	Accuracy	MCC
AA + PSSM	76.8	97	94.2	0.75
AA + SS	78	97	94.4	0.76
SS + PSSM	78	97	94.4	0.76
AA + SS + DIPEP + PSSM	76.8	96.8	94.1	0.74
Cascade module	78	97.3	94.7	0.77

AA, amino acid composition; *PSSM*, position-specific scoring matrices; *SS*, secondary structure composition; *DIPEP*, dipeptide composition

Cascade SVM module

Machine learning techniques are often incapable of handling the noise produced due to the large number/complex nature of the input units/patterns, which therefore affects their classification efficiencies. This can be exemplified for the systems developed using hybrid modules that consist of PSSM profiles and dipeptide compositions. However, one solution to this inadequacy is to construct a cascade SVM module. The present cascade module consisted of two layers. A brief description of these layers is given below.

First layer

In the first layer, four individual modules based on protein features such as amino acid compositions, dipeptide composition, PSSM compositions and secondary structure composition were developed.

Second layer

The second layer received the prediction results in the form of SVM-predicted scores provided by the individual trained SVM models constructed in the first layer to train the second layer SVM model. Here, SVM was provided with a vector of four dimensions (one for amino acid composition, one for dipeptide composition, one for PSSM and one for secondary structure-based results). Hence, the second layer correlates the predicted information from the first layer models to provide the final output. The SVM classifier thus generated provides the predictions to discriminate whether the proteins are ordered or disordered.

The results obtained were quite promising, with improved sensitivity and MCC values obtained. Using the RBF kernel ($g = 4$, $c = 15$, $j = 2$), the sensitivity, specificity, accuracy and MCC values were found to be 78, 97.3, 94.7, and 0.77 (Table 2), respectively.

PSSM-based SVM modules generated using different databases

All of the SVM modules developed so far were observed to be inadequate for improving the performance accuracy of the classifier. Efforts to further improve the efficiency of the classifier were therefore carried out. The next approach we exploited was the generation of PSSM profiles against different datasets, such as PDB, SWISS-PROT, etc. The results obtained using the SVM modules for PSSM profiles generated against the NR database were discussed in the sections above. Next, the training of the SVM model by PSSM profiles generated using the SWISS-PROT database yielded sensitivity, specificity, accuracy and MCC values of 78, 95.4, 93.1% and 0.71, respectively, which was observed to be a poorer performance than that given by the NR database-generated PSSM profile composition-based SVM module. Further, the use of PSSMs generated from the PSI-BLAST search against PDB database as an input feature achieved a sensitivity of 78% with a corresponding increase in specificity, accuracy and MCC; the values were 97.7, 95.1%, 0.78, respectively ($g = 5$, $c = 3$, $j = 3$), as shown in Table 3. Hence, much improved results were obtained when the PSSM profiles generated against the PDB database were used in comparison with the other databases. The model thus generated was therefore used as a classifier for the prediction of disordered proteins. Table 3 shows the detailed performances of SVM modules trained on PSSM profiles generated from different databases.

Table 3 Performances of PSSM (generated using different databases) composition-based SVM modules in predicting disordered proteins

Database	Sensitivity	Specificity	Accuracy	MCC
NR	76.8	97.0	94.2	0.75
SWISS-PROT	78	95.4	93.1	0.71
PDB	78	97.7	95.1	0.78

NR, nonredundant database; *PDB*, protein data bank

Evaluation on partially disordered proteins

Quite often disorderedness is not a function of the complete protein, but rather of some regions in the protein. Such proteins are said to be partially disordered. Our method yielded a good performance for the prediction of such proteins. Independent testing of the approach was done on a dataset of 417 partially disordered proteins. It was observed that out of 417 proteins, 86.63% proteins were predicted to be disordered. Our method is thus quite sensitive when applied to the prediction of partially disordered proteins compared to the previously proposed method, POODLE, which could detect only 16.67% of the moderately (20–40%) disordered proteins.

Implementation of the DPROT server

Our work in this study led to the development of a public web server, "DPROT," that can be used for the prediction of disordered proteins. DPROT is available at http://www.imtech.res.in/raghava/dprot/. This is a user-friendly server developed on a SUN server running under the Solaris environment using HTML, PERL and CGI-PERL. The user may paste or upload their protein sequence in standard FASTA format. The server also allows the user to specify the threshold. The web server automatically generates the

PSSM of the given sequence by carrying out the PSI-BLAST search against the PDB sequence database and uses the profile composition as an input to a trained SVM model for prediction. Information regarding whether the query protein is ordered or disordered is displayed along with the SVM predicted score.

Conclusions

To conclude, the present work is an attempt to improve the performance of disordered protein prediction using evolutionary information hidden in a PSSM profile. Initial attempts focused on the use of individual modules, such as for amino acid composition, dipeptide composition, etc. Later, hybrid modules were tested for their performance. Part of the study also focused on the development of cascade SVM-based modules. Unfortunately, a significant increase in prediction accuracy was not obtained. PSSM profiles were generated on different databases and we were able to model an efficient SVM classifier from this information. A server called DPROT was developed based on the results obtained.

Acknowledgments The authors are thankful to the Council of Scientific and Industrial Research (CSIR) and the Department of Biotechnology, Government of India for financial assistance.

References

Bhasin M, Raghava GPS (2004), ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. Nucleic Acids Res 32:414–419

Chou KC, Shen HB (2007a) MemType-2L: a web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. Biochem Biophys Res Comm 360:339–345

Chou KC, Shen HB (2007b) Recent progresses in protein subcellular location prediction. Anal Biochem 370:1–16

Dosztanyi Z, Csizmok V, Tompa P, Simon I (2005) The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. J Mol Biol 347:827–839

Dunker AK, Obradovic Z (2001) The protein trinity-linking function and disorder. Nat Biotechnol 19:805–806

Dunker AK, Brown CJ, Obradovic Z (2002) Identification and functions of usefully disordered proteins. Adv Protein Chem 62:25–49

Fink AL (2005) Natively unfolded proteins. Curr Opin Struct Biol 15:35–41

Galzitskaya OV, Garbuzynskiy SO, Lobanov MY (2006) FoldUnfold: web server for the prediction of disordered regions in protein chain. Bioinformatics 22:2948–2949

Garg A, Bhasin M, Raghava GPS (2005a) Support vector machine-based method for subcellular localization of human proteins using amino acid compositions, their order, and similarity search. J Biol Chem 280:14427–14432

Garg A, Kaur H, Raghava GPS (2005b) Real value prediction of solvent accessibility in proteins using multiple sequence alignment and secondary structure. Proteins 61:318–325

Joachims T (1999) Making large-scale SVM learning particle. In: Scholkopf B, Burges C, Smola A (eds) Advances in kernel methods support vector learning. MIT Press, Cambridge, MA, pp 42–56

Jones DT (1999) Protein secondary structure prediction based on position specific scoring matrices. J Mol Biol 292:195–202

Jones DT, Ward JJ (2003) Prediction of disordered regions in proteins from position specific score matrices. Proteins 53:573–578

Kaur H, Raghava GPS (2002) BetaTPred: Prediction of beta turns in a protein using statistical algorithms. Bioinformatics 18:498–499

Kaur H, Raghava GPS (2003) A neural-network based method for prediction of gamma-turns in proteins from multiple sequence alignment. Protein Sci 2:923–929

Kaur H, Raghava GPS (2004a) Prediction of alpha-turns in proteins using PSI-BLAST profiles and secondary structure information. Proteins 55:83–90

Kaur H, Raghava GPS (2004b) A neural network method for prediction of β-turn types in proteins using evolutionary information. Bioinformatics 20:2751–2758

Kumar M, Bhasin M, Natt NK, Raghava GPS (2005) BhairPred: prediction of b-hairpins in a protein from multiple alignment information using ANN and SVM techniques. Nucleic Acids Res 33:154–159

Lata S, Sharma BK, Raghava GPS (2007) Analysis and prediction of antibacterial peptides. BMC Bioinformatics 8:263

Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB (2003a) Protein disorder prediction: implications for structural proteomics. Structure 11:1453–1459

Linding R, Russell RB, Neduva V, Gibson TJ (2003b) GlobPlot: exploring protein sequences for globularity and disorder. Nucleic Acids Res 31:3701–3708

Radivojac P, Obradovic Z, Smith DK, Zhu G, Vucetic S, Brown CJ, Lawson JD, Dunker AK. (2004) Protein flexibility and intrinsic disorder. Protein Sci 13:71–80

Rashid M, Saha S, Raghava GPS (2007) Support vector machine-based method for predicting subcellular localization of mycobacterial proteins using evolutionary information and motifs. BMC Bioinformatics 8:337

Romero P, Obradovic Z, Dunker AK (1999) Folding minimal sequences: the lower bound for sequence complexity of globular proteins. FEBS Lett 462:363–367

Romero P, Obradovic Z, Li X, Garner EC, Brown CJ, Dunker AK (2001) Sequence complexity of disordered protein. Proteins 42:38–48

Shen HB, Chou KC (2007a) EzyPred: a top-down approach for predicting enzyme functional classes and subclasses. Biochem Biophys Res Comm 364:53–59

Shen HB, Chou KC (2007b) Nuc-PLoc: a new web-server for predicting protein subnuclear localization by fusing PseAA composition and PsePSSM. Protein Eng Des Sel 20:561–567

Shimizu K, Muraoka Y, Hirose S, Tomii K, Noguchi T (2007) Predicting mostly disordered proteins by using structure-unknown protein data. BMC Bioinformatics 8:78

Sussman JL, Prilusky J, Felder CE, Zeev-Ben-Mordehai T, Rydberg EH, Man O, Beckmann JS, Silman I (2005) Fold index: a simple tool to predict whether a given protein sequence is intrinsically unfolded. Bioinformatics 21:3435–3438

Vihinen M, Torkkila E, Riikonen P (1994) Accuracy of protein flexibility predictions. Proteins 19:141–149

Vucetic S, Obradovic Z, Vacic V, Radivojac P, Peng K, Iakoucheva LM, Cortese MS, Lawson JD, Brown CJ, Sikes JG, Newton CD, Dunker AK (2005) DisProt: a database of protein disorder. Bioinformatics 21:137–140

Ward JJ, McGuffin LJ, Bryson K, Buxton BF, Jones DT (2004) The DISOPRED server for the prediction of protein disorder. Bioinformatics 20:2138–2139

Weinreb PH, Zhen W, Poon AW, Conway KA, Lansbury PT Jr (1996) NACP, a protein implicated in Alzheimer's disease and learning, is natively unfolded. Biochemistry 35:13709–13715

Wright PE, Dyson HJ (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. J Mol Biol 293:321–331

Xie D, Li A, Wang M, Fan Z Feng H (2005) LOCSVMPSI: a web server for subcellular localization of eukaryotic proteins using SVM and profile of PSI-BLAST. Nucleic Acids Res 33:105–110

Xie Q, Arnold GE, Romero P, Obradovic Z, Garner E, Dunker AK (1998) The sequence attribute method for determining relationships between sequence and protein disorder. Genome Inform 9:193–200

Yang ZR, Thomson R, McNeil P, Esnouf RM (2005) RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. Bioinformatics 21:3369–3376

Yip YL, Scheib H, Diemand AV, Gattiker A, Famiglietti LM, Gasteiger E, Bairoch A (2004) The Swiss-Prot variant page and the ModSNP database: a resource for sequence and structure information on human protein variants. Hum Mutat 23:464–470

Zhi-Qiang Ye, Zhao SQ, Gao G, Liu XQ, Langlois RE, Lu H, Wei L (2007) Finding new structural and sequence attributes to predict possible disease association of single amino acid polymorphism (SAP). Bioinformatics 23:1444–1450