Journal of Theoretical Biology 254 (2008) 799–803

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Using Bayesian multinomial classifier to predict whether a given protein sequence is intrinsically disordered

Alla Bulashevska [a,*], Roland Eils [a,b]

[a] Department of Theoretical Bioinformatics, German Cancer Research Center, Im Neuenheimer Feld 280, 69120 Heidelberg, Germany
[b] Department of Bioinformatics and Functional Genomics, Institute of Pharmacy and Molecular Biotechnology (IPMB), University of Heidelberg, Germany

ARTICLE INFO

Article history:
Received 19 November 2007
Received in revised form
19 May 2008
Accepted 19 May 2008
Available online 14 June 2008

Keywords:
Unfolded proteins
Disorder prediction
Model-based classification
Multinomial model

ABSTRACT

Intrinsically disordered proteins (IDPs) lack a well-defined three-dimensional structure under physiological conditions. Intrinsic disorder is a common phenomenon, particularly in multicellular eukaryotes, and is responsible for important protein functions including regulation and signaling. Many disease-related proteins are likely to be intrinsically disordered or to have disordered regions. In this paper, a new predictor model based on the Bayesian classification methodology is introduced to predict for a given protein or protein region if it is intrinsically disordered or ordered using only its primary sequence. The method allows to incorporate length-dependent amino acid compositional differences of disordered regions by including separate statistical representations for short, middle and long disordered regions. The predictor was trained on the constructed data set of protein regions with known structural properties. In a Jack-knife test, the predictor achieved the sensitivity of 89.2% for disordered and 81.4% for ordered regions. Our method outperformed several reported predictors when evaluated on the previously published data set of Prilusky et al. [2005. FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. Bioinformatics 21 (16), 3435–3438]. Further strength of our approach is the ease of implementation.

1. Introduction

It is well established that proteins fold to their unique native conformations as determined by their amino acid sequences (Anfinsen, 1973). However, there are proteins that are unable to maintain well-defined structures even under physiological conditions. These proteins are often called natively unfolded or intrinsically disordered proteins (IDPs) and assume either partially folded or completely unfolded conformations.

IDPs are involved in numerous processes in the cell: transcriptional activation, cell-cycle regulation, membrane transport, and signalling (Wright and Dyson, 1999). Twenty-eight different functions were found to be associated with disordered regions (Dunker et al., 2002), which can be grouped into four broad classes: molecular recognition, molecular assembly and/or disassembly, providing sites for protein post-translational modification and entropic chain activities.

The common occurrence of intrinsic disorder in cancer-associated and signaling proteins (Iakoucheva, 2002) suggests their potential involvement in the pathogenesis of cancer. IDPs also play key roles in diseases mediated by protein misfolding and aggregation (Bates, 2003; Kaplan et al., 2003), in cardiovascular (Cheng et al., 2006) and autoimmune (Carl et al., 2005) diseases.

The series of papers dedicated to the functional anthology of intrinsic disorder has been recently published (Xie et al., 2007a, b; Vucetic et al., 2007).

A database for the deposition of disordered protein information was developed (http://www.disprot.org), which contains 469 proteins with 1114 regions experimentally characterized as disordered (release: 3.5, 12/22/2006). Since the experimental methods to study protein disorder such as nuclear magnetic resonance (NMR) and circular dichroism (CD), Raman spectroscopy or X-ray crystallography are expensive, there is a crucial need for novel bioinformatics approaches that allow to learn from a few experimentally verified examples and to predict fold property for much larger groups of known and potential proteins. Computer-based scanning of the protein sequence for potentially disordered regions can also assist target selection process for high-throughput protein structure determination by removing disordered targets (Oldfield et al., 2005).

Various computational approaches have been already developed to predict protein disorder (see the list of references at http://www.disprot.org/predictors.php). Nearly half of them are based on the neural network learning model; some predictors employ support vector machines. It was shown in several studies that the primary structure of disordered regions is distinct from

---

[*] Corresponding author. Tel.: +49 6221 423605.
E-mail address: A.Bulashevska@dkfz.de (A. Bulashevska).

0022-5193/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2008.05.040

that of structured regions. Disordered regions usually have low sequence complexity (Romero et al., 2001) and are enriched in charged or polar residues and depleted in hydrophobic residues (Hansen et al., 2006). Several prediction methods are based on the amino acid composition (AAC) of the protein sequence, e.g. VL3 of Obradovic et al. (2003) and DISOPRED (Ward et al., 2004). PONDR (Romero et al., 2001) considers both AAC and physicochemical property based attributes, including aromaticity, net charge, flexibility, and hydropathy. DISOPRED applies 21 input parameters per residue. The features are usually extracted from the partial amino acid sequence within a sliding window and a binary classifier is then built to predict the probability of a residue being in a disordered or ordered region. The per-residue disorder predictors, such as VL3, GlobPlot (Linding et al., 2003) and DISOPRED, require a fixed window size to be chosen. Different widths for the sequence window usually have to be tried.

Various predictors adopt different definitions of disorder. For example, NORSp (Liu and Rost, 2003) focuses on long regions having no regular secondary structure, which were not considered to be disordered by Vucetic et al. (2005). Some predictors are specifically designed to predict short disordered regions, while others are tailored for long disordered regions (>30 residues). The VSL2 predictor of Peng et al. (2006) models short and long disordered regions separately and utilizes a meta predictor to integrate the specialized predictors into the final predictor model.

Several predictors have been developed for predicting whether or not given proteins or protein regions/domains are intrinsically disordered (Han et al., 2005; Shimizu et al., 2007; Weathers et al., 2004). IUPred (Dostanyi et al., 2005a, b) uses the pairwise energy content estimated from AAC to distinguish between folded and unfolded proteins/regions. RONN (Yang, 2005) aligns a query protein sequence to a set of prototype disordered/ordered regions and uses the alignment scores to classify the query sequence. FoldIndex of Prilusky et al. (2005) predicts whether a given protein sequence is intrinsically unfolded by plotting the average hydrophobicity of the residues in the sequence against the net charge of the sequence as was proposed by Uversky et al. (2000). Garbuzynskiy et al. (2004) proposed plotting of the average hydrophobicity against the number of residues in contact.

In this paper, we introduce a new prediction method, which exploits the Bayesian classification procedure to predict disordered property for a given protein or protein region from its primary sequence. Bayesian Markov chain model-based classification has already found its application in proteomics for the prediction of protein subcellular locations (Bulashevska and Eils, 2006). This approach represents each class with a single probabilistic summary. Since the AAC of disordered regions is distinct from that of ordered, we propose to use multinomial models for the description of class-conditional densities. The intuition behind this approach is that each protein sequence belonging to a certain class can be considered as a realization of an independent random process that emits symbols from an alphabet of 20 amino acids.

Peng et al. (2006) compared the amino acid composition of short (4–30 residues) and long (>30 residues) disordered regions to the composition of a reference ordered data set Globular-3D. Both types of disordered regions exhibited similar overall compositional bias that characterizes intrinsic protein disorder, i.e. depletion of the typically buried W, C, F, I, Y, V, L and enrichment of the typically exposed K, E, P, S, Q, R. However, some significant differences were also found. Short disordered regions are more depleted in C, I, V and L, while long disordered regions are more enriched in K, E and P but are less enriched in Q and S. In addition, long disordered regions are depleted in G and N, while short disordered regions are enriched in G and D. In order to incorporate length-dependent properties of disordered regions, we model short ($\leqslant 30$ residues), middle (31–100 residues) and long (>100 residues) disordered regions separately.

The parameters of the multinomial models corresponding to the four classes (short, middle, long disordered and ordered) were estimated from the constructed training data set of disordered and ordered regions.

In a Jack-knife test, our prediction method achieved the predictive accuracies of 89.2% and 81.4% for disordered and ordered regions, respectively.

2. Materials and methods

2.1. Training data set

The sequences of disordered regions for training were extracted from DisProt (Release 3.5.2006) (Vucetic et al., 2005); 1077 disordered regions were extracted. Length distribution of the disordered regions and the corresponding number of residues is shown in Table 1.

The ordered sequences were extracted from PDB-SELECT-25 (2006 version) (Hobohm and Sander, 1994). The PDB-SELECT database is a subset of the structures in the PDB that does not contain (highly) homolog sequences. PDB-SELECT-25 shows less than 25% sequence homology. From the PDB-SELECT-25 sequences of 709 regions from higher resolution crystal structures ($<2A$) with no missing backbone or side chain coordinates and no non-standard amino acid residues were selected.

2.2. Multinomial models

Multinomial models assume a *bag-of-amino acid* sequence representation, which considers the appearance of each amino acid as an independent event. The order in which amino acids occur in a given amino acid sequence is ignored; the only information retained is a vector of counts $\mathbf{n} = (n_1, \ldots, n_{20})$, where n_i is the number of occurrences of amino acid i in the sequence.

We assume that the probability of a sequence s to come from a certain class c is given by a multinomial probability function governed by its vector of parameters $\theta_c = (\theta_{c1}, \ldots, \theta_{c20}) \in [0,1]^{20}$:

$$p(s|\theta_c) = \frac{n!}{\prod_{i=1}^{20} n_i!} \prod_{i=1}^{20} \theta_{ci}^{n_i}, \qquad (1)$$

where $n = \sum_i n_i$ denotes the length of the sequence. The parameter θ_{ci} denotes the cth class-conditional probability of amino acid i to occur in a sequence. The parameters of the model corresponding to class c are estimated from the training regions belonging to the class c. Thus, the parameter θ_{ci} is calculated as

$$\theta_{ci} = \frac{n_{ci}}{\sum_{i=1}^{20} n_{ci}}, \qquad (2)$$

where n_{ci} is the number of occurrences of amino acid i in the sequences of class c. This way of estimating parameters of the

Table 1
Number of regions in our training data set for each of the four classes modeled by the Bayesian multinomial classifier

Class	# Regions
Short ($\leqslant 30$)	683
Middle (31–100)	247
Long (>100)	147
Total disordered	1077
Ordered	709
Total	1786

model is called *maximum likelihood estimation*, because it can be shown that using the frequencies to calculate the probabilities maximizes the total probability of training instances given the model (the *likelihood*). For a detailed description of multinomial models we refer the reader to the book by Durbin et al. (1998).

2.3. Bayesian multinomial classifier

Bayesian classification is a widely applied method in the machine learning and statistical community, which is based on Bayes' theorem (Bayes rule). According to Bayes' rule, the class for an unlabeled sequence s can be inferred using the posterior probability:

$$p(c|s) = \frac{p(c)p(s|c)}{p(s)} = \frac{p(c)p(s|c)}{\sum_c p(c)p(s|c)}. \tag{3}$$

We assume class prior probabilities $p(c)$ to be equally distributed. We further assume that the sequences of each class are generated from multinomial models. Thus, given the parameters $\{\theta_c\}$ of the models for each class, the term $p(s|c)$ denoting the prior probability of a sequence s to belong to the class c can be computed using the formula (1) for $p(s|\theta_c)$ from previous subsection.

Since we model short, middle and long disordered regions separately, the estimation of the class-conditional densities involves four subproblems (for short, middle, long disordered and ordered classes), in which each of the class-conditional density is estimated based on the data belonging to the corresponding class only.

Bayesian classifier is a *probabilistic* classifier, which yields for each query instance the posterior probability for each class, a numeric value that represents the degree to which an instance is a member of a class. To produce a discrete output, the following decision rule is usually applied: the class should be the one which maximizes the posterior probability.

To classify an input sequence as disordered or ordered, we sum the posterior probabilities for short, middle and long disordered subtypes into a single value describing the posterior probability of a sequence to be disordered and then use the standard decision rule to come up with a discrete output, i.e. predict one of the two classes (disordered/ordered) showing the biggest posterior probability.

2.4. Performance evaluation

The prediction performance of our predictor was validated with *Jack-knife test* (or *leave-one-out cross-validation*) (Mardia et al., 1979). By Jack-knife test the learning step is performed with all training instances except the one for which the class is to be predicted.

The prediction quality was evaluated using the standard measures of *sensitivity* (SN) and *specificity* (SP), where the sensitivity, or *true positive rate*, is the percentage of disordered sequences correctly predicted, and the SP, or *true negative rate*, is the percentage of ordered sequences correctly predicted. We calculate the *overall accuracy* (ACC) as the average of SN and SP, which is more suitable than the percentage of all correctly predicted sequences for data sets with imbalanced class distributions. We also show receiver operating characteristic (ROC) curve and report area under the ROC curve (AUC) calculated using the R package ROCR (Sing et al., 2005).

3. Results and discussion

The confusion matrix of the prediction results of our predictor is given in Table 2.

Table 2
Confusion matrix of the results of our predictor

	Predicted group		
	Disordered	Ordered	Sum
Disordered	**961**	116	1077
Ordered	132	**577**	709
Sum	1093	693	1786

Bold numbers along the major diagonal represent the numbers of correctly predicted sequences for each class, the numbers off this diagonal represent the errors.

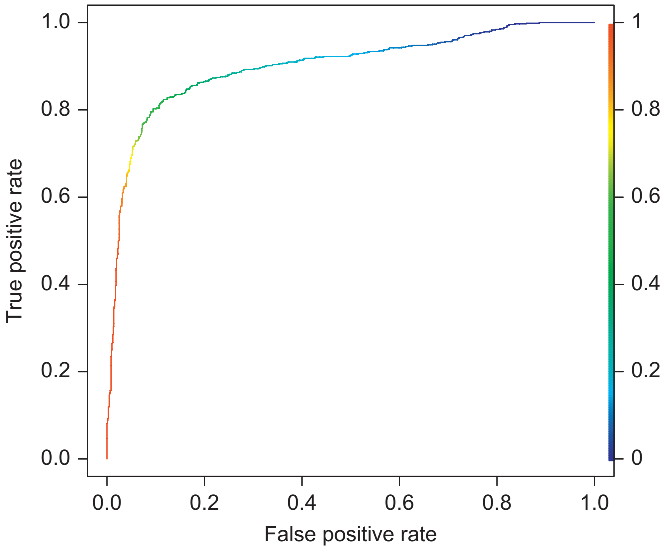

Fig. 1. ROC curve.

Table 3
Prediction accuracies of our predictor (see Performance evaluation)

SN	SP	ACC	AUC
89.2	81.4	85.3	0.9

Fig. 1 shows the ROC curve constructed using the probabilistic outputs of our Bayesian multinomial classifier. The corresponding AUC value achieved was 0.9.

Table 3 summarizes the prediction accuracies achieved with our predictor.

It is remarkable that for disordered regions our method reaches the SN of 89.2%. This excellent performance of our predictor concerning the sensitivity for disorder might be explained with the potential of the multinomial modelling of length-dependent subtypes of disordered regions. The high SN was achieved while obtaining the high SP of 81.4%.

Compared with the semi-supervised spectral graph partitioning method of Shimizu et al. (2007), which has achieved the SN of 72.3% and the SP of 97.7%, our method shows the more balanced SN and SP, though the ACC of both methods is almost the same (around 85%).

Table 4 demonstrates the confusion matrix and the prediction accuracies for each of the three subtypes of disordered regions. The results reported in Table 4 suggest that short disordered regions can be even more accurately discriminated with our predictor than middle or long ones.

Table 4
Prediction results for three subtypes of disordered regions

	Predicted group			
	Disordered	Ordered	Sum	Accuracy
Short	**627**	56	683	91.8
Middle	**215**	32	247	87.0
Long	**119**	28	147	81.0

Bold numbers indicate the short, middle and long disordered regions correctly predicted as disordered.

Table 5
Comparison of the prediction accuracies of our predictor and four other algorithms obtained for the data set of Prilusky et al. (2005)

	SN	SP	ACC
FoldIndex	76.9	88.1	82.5
	30/39	133/151	
DISOPRED	56.4	98.7	77.6
	22/39	149/151	
PONDR	71.8	92.7	82.3
	28/39	140/151	
GlobPlot	23.1	98.0	60.6
	9/39	148/151	
Our predictor	89.7	89.4	89.6
	35/39	135/151	

To independently evaluate our predictor, we tested it on the data set of 39 intrinsically unfolded and 151 folded proteins (or domains) compiled by Prilusky et al. (2005). In Prilusky et al. (2005) the results of FoldIndex predictor on this data set were reported and compared with the results of three other predictors (DISOPRED, PONDR and GlobPlot). In order to compare the methods a fold score for the entire sequence was obtained from the scores of the individual residues by calculating the arithmetic (for PONDR and GlobPlot) and geometric (for DISOPRED) means. FoldIndex achieved the SN of 76.9%, which was the highest among all four predictors. DISOPRED, PONDR and GlobPlot were very good in correct identification of ordered regions, but their sensitivity for disorder was less impressive. Table 5 compares the prediction accuracies achieved with our predictor with the results reported in Prilusky et al. (2005). For disordered sequences our method reaches the remarkable SN of 89.7%. Our method also shows the well balanced SN and SP. None of the four predictors reaches our ACC of 89.6%.

4. Conclusion

We introduced a new approach for the prediction of fold property for a given protein or protein region based on its primary sequence information alone.

The employment of multinomial models for the description of class-conditional distributions allows one to make better use of sequence AAC, which is an important feature previously adopted to discriminate between disordered and ordered sequences. The Bayesian multinomial classifier contains multiple probabilistic summaries for the disordered regions of different lengths, which provide the opportunity of better representing length-dependent compositional differences. Our predictor is applicable to the regions of any length.

Our predictor achieved high prediction accuracies and outperforms several previously reported predictors.

The method can be effectively implemented and is computationally efficient.

We hope that our prediction method will provide support for relating disorder to protein function and help to translate the new discoveries into new druggable targets.

Acknowledgement

This work was supported by the National Genome Research Network (NGFN) funded by BMBF 01GR0450.

References

Anfinsen, C.B., 1973. Principles that govern the folding of protein chains. Science 181, 223–230.

Bates, G., 2003. Huntington aggregation and toxicity in Huntington's disease. Lancet 361, 1642–1644 (review).

Bulashevska, A., Eils, R., 2006. Predicting protein subcellular locations using hierarchical ensemble of Bayesian classifiers based on Markov chains. BMC Bioinformatics 71, 298.

Carl, P.L., Temple, B.R., Cohen, P.L., 2005. Most nuclear systemic autoantigens are extremely disordered proteins: implications for the etiology of systemic autoimmunity. Arthritis Res. Ther. 7, R1360–R1374.

Cheng, Y., LeGall, T., Oldfield, C.J., Dunker, A.K., Uversky, V.N., 2006. Abundance of intrinsic disorder in protein associated with cardiovascular disease. Biochemistry 45 (35), 10448–10460.

Dosztanyi, Z., Csozmok, V., Tompa, P., Simon, I., 2005a. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics 21 (16), 3433–3434.

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

Dunker, A.K., Brown, C.J., Lawson, J.D., Obradovic, I.M., 2002. Intrinsic disorder and protein function. Biochemistry 41 (21), 6573–6582.

Durbin, R., Eddy, S., Krogh, A., Mitchison, G., 1998. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press, Cambridge.

Garbuzynskiy, S.O., Lobanov, M.Y., Galzitskaya, O.V., 2004. To be folded or to be unfolded? Protein Sci. 13, 2871–2877.

Han, P., Zhang X., Feng Z.P., Norton R.S., 2005. Predicting intrinsically unstructured proteins based on amino acid composition. Proc. AusDM05, 131–140.

Hansen, J.C., Lu, X., Ross, E.D., Woody, R.W., 2006. Intrinsic protein disorder, amino acid composition, and histone terminal domains. J. Biol. Chem. 281, 1853–1856.

Hobohm, U., Sander, C., 1994. Enlarged representative set of protein structures. Protein Sci. 33, 522–524.

Iakoucheva, L.M., 2002. Intrinsic disorder in cell-signaling and cancer-associated proteins. J. Mol. Biol. 323, 573–584.

Kaplan, B., Ratner, V., Haas, E., 2003. Alpha-synuclein: its biological function and role in neurodegenerative diseases. J. Mol. Neurosci. 20, 83–92 (review).

Linding, R., Russell, R.B., Neduva, V., Gibson, T.J., 2003. GlobPlot: exploring protein sequences for globularity and disorder. Nucleic Acids Res. 31, 3701–3708.

Liu, J., Rost, B., 2003. NORSp: predictions of long regions without regular secondary structure. Nucleic Acids Res. 31, 3833–3835.

Mardia, K.V., Kent, J.T., Bibby, J.M., 1979. Multivariate Analysis. Academic Press, London.

Obradovic, Z., Peng, K., Vucetic, S., Radivojac, P., Brown, C.J., Dunker, A.K., 2003. Predicting intrinsic disorder from amino acid sequence. Proteins Struct. Funct. Bioinf. 536, 566–572.

Oldfield, C.J., Ulrich, E.L., Cheng, Y., Dunker, A.K., Markley, J.L., 2005. Addressing the intrinsic disorder bottleneck in structural proteomics. Proteins: Struct. Funct. Bioinf. 59, 444–453.

Peng, K., Radivojac, P., Vucetic, S., Dunker, A.K., Obradovic, Z., 2006. Length-dependent prediction of protein intrinsic disorder. BMC Bioinformatics 7, 208.

Prilusky, J., Felder, C.E., Zeev-Ben-Mordehai, T., Rydberg, E.H., Man, O., Beckmann, J.S., Silman, I., Sussman, J.L., 2005. FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. Bioinformatics 21 (16), 3435–3438.

Romero, P., Obradovic, Z., Li, X., Garner, E.C., Brown, C.J., Dunker, A.K., 2001. Sequence complexity of disordered protein. Proteins 421, 38–48.

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.

Sing, T., Sander, O., Beerenwinkel, N., Lengauer, T., 2005. ROCR: visualizing classifier performance in R. Bioinformatics 21, 3940–3941.

Uversky, V.N., Gillespie, J.R., Fink, A.L., 2000. Why are "natively unfolded" proteins unstructured under physiologic conditions? Proteins 41, 415–427.

Vucetic, S., Obradovic, Z., Vacic, V., Radivojac, P., Peng, K., Iakoucheva, L.M., Cortese, M.S., Lawson, J.D., Brown, C.J., Sikes, J.G., Newton, C.D., Dunker, A.K., 2005. DisProt: a database of protein disorder. Bioinformatics 21, 137–140.

Vucetic, S., Xie, H., Iakoucheva, L.M., Oldfield, C.J., Dunker, A.K., Obradovic, Z., Uversky, V.N., 2007. Functional anthology of intrinsic disorder. 2. Cellular components, domains, technical terms, developmental processes, and coding sequence diversities correlated with long disordered regions. J. Proteome Res. 6 (5), 1899–1916.

Ward, J.J., McGuffin, L.J., Bryson, K., Buxton, B.F., Jones, D.T., 2004. The DISOPRED server for the prediction of protein disorder. Bioinformatics 20, 2138–2139.

Weathers, E.A., Paulaitis, M.E., Woolf, T.B., Hoh, J.H., 2004. Reduced amino acid alphabet is sufficient to accurately recognize intrinsically disordered protein. FEBs Lett. 576, 348–352.

Wright, P.E., Dyson, H.J., 1999. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. J. Mol. Biol. 293, 321–331.

Xie, H., Vucetic, S., Iakoucheva, L.M., Oldfield, C.J., Dunker, A.K., Uversky, V.N., Obradovic, Z., 2007a. Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions. J. Proteome Res. 6 (5), 1882–1898.

Xie, H., Vucetic, S., Iakoucheva, L.M., Oldfield, C.J., Dunker, A.K., Obradovic, Z., Uversky, V.N., 2007. Functional anthology of intrinsic disorder. 3. Ligands, post-translational modifications, and diseases associated with intrinsically disordered proteins. J. Proteome Res. 6 (5), 1917–1932.

Yang, Z.R., 2005. RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. Bioinformatics 21 (16), 3369–3376.