Osong Public Health Res Perspect 2014 5(4), 211–218
http://dx.doi.org/10.1016/j.phrp.2014.06.006
pISSN 2210-9099 eISSN 2233-6052

ORIGINAL ARTICLE

A Novel Approach for Predicting Disordered Regions in A Protein Sequence

Meijing Li [a], Seong Beom Cho [b], Keun Ho Ryu [a,*]

[a]*Database/Bioinformatics Laboratory, Chungbuk National University, Cheongju, Korea.*
[b]*Division of Bio-Medical Informatics, Center for Genome Science, Korea National Institute of Health, Cheongju, Korea.*

Received: June 19, 2014
Revised: June 24, 2014
Accepted: June 24, 2014

KEYWORDS:
amino acid sequence,
disordered protein,
emerging subsequence,
protein structure

Abstract
Objectives: A number of published predictors are based on various algorithms and disordered protein sequence properties. Although many predictors have been published, the study of protein disordered region prediction is ongoing because different prediction methods can find different disordered regions in a protein sequence.
Methods: Therefore we have used a new approach to find the more varying disordered regions for more efficient and accurate prediction of protein structures. In this study, we propose a novel approach called "emerging subsequence (ES) mining" without using the characteristics of the disordered protein. We first adapted the approach to generate emerging protein subsequences on public protein sequence data. Second, the disordered and ordered regions in a protein sequence were predicted by searching the generated emerging protein subsequence with a sliding window, which tends to overlap. Third, the scores of the overlapping regions were calculated based on support and growthrate values in both classes. Finally, the score of predicted regions in the target class were compared with the score of the source class, and the class having a higher score was selected.
Results: In this experiment, disordered sequence data and ordered sequence data was extracted from DisProt 6.02 and PDB respectively and used as training data. The test data come from CASP 9 and CASP 10 where disordered and ordered regions are known.
Conclusion: Comparing with several published predictors, the results of the experiment show higher accuracy rates than with other existing methods.

1. Introduction

The study of protein structure for the prediction of function using data mining has always been known as an important research topic in Bioinformatics. Disordered proteins, referred to as naturally unfolded proteins or intrinsically unstructured proteins, are characterized by a lack of stable tertiary structure when the protein exists as an isolated polypeptide chain under physiological conditions *in vitro*. However, all the analyses of protein

*Corresponding author.
E-mail: khryu@dblab.chungbuk.ac.kr

are based on protein primary structure denoted amino acid sequences. Protein sequences decide protein structure, and protein structures concern protein function. In the study of protein structures, prediction of disordered regions in a protein sequence is an important topic [1]. The reasons are as follows: (1) Proteins can function when protein disordered sequences fold with other protein sequences. Therefore, finding the protein disordered regions helps to study functions of proteins [2]. Moreover, most of the hub proteins cannot highly interact with proteins compared with nonhub proteins [3] (disordered proteins) except cancer proteins [4]. (2) When we analyzed the similarity between proteins by protein alignments, identification of disordered regions could avoid disordered regions compared with ordered regions, which therefore improved the accuracy of analysis. (3) Eukaryotic Linear Motifs (ELMs) which are short linear peptide regions containing independent functions not related to protein structures. However, the 70% of ELMs are located in disordered regions [5]. (4) In sequence data, division between disordered regions and ordered regions are more beneficial to study three-dimensional protein structures and properties from protein sequences [6].

In early 1997, Romero et al. proposed the first protein disordered region predictor which applied data mining algorithms to protein sequence data without fixed protein three-dimensional structures [7]. To date, a number of predictors of protein disordered regions have been published. From a view of algorithms which were used to construct the prediction model, several data mining and machine learning algorithms were applied, such as nearest neighbor algorithm [8], support vector machines (SVMs) [9−14], neural networks (NNs) [15−23], artificial neural network (ANNs), regression [24−26], sliding window [27,28], random forest [29], Bayesian Markov chain model [30] and so on.

Many protein properties were used to study protein disordered regions, for example low hydrophobicity, the content of B-factor (residues with high B-factor loops) [31], position-specific score [32−35], high net charge and low hydrophobicity [27], low contact density (average amino acid contact propensity scores with or without pairwise interaction energy matrices) [37−39] and so on. Recently, two predictors [29] were proposed which are based on the profiles of amino acid indices representing various physiochemical and biochemical properties of the 20 amino acids. DISOclust [40] used a different method from other methods which was based on the analysis of three-dimensional structural models using ModFOLDclust [41].

In addition, to increase the accuracy of prediction, several meta-predictors were developed which were combined with several predictors [10,18,21,42−46]. Apart from these methods, multiple sequence alignment with proteins of known protein domains is used to analyze protein structures.

Although many meta-predictors are proposed for increasing prediction accuracy, the increase of accuracy is limited to published based models. We also need to propose new basic prediction methods to search the disordered regions which have specific characteristics using different methods. According to the characteristics of disordered proteins, the regions which are predicted are different from each other [47]. Most of the protein disordered region predictors applied characteristics of disordered proteins to identify the disordered region in a protein sequence. In this study, a novel approach was proposed which did not apply the characteristics of disordered protein. In this paper, we modified and applied an emerging substring generation algorithm which was based on a suffix tree to derive the protein emerging subsequences [36]. These protein emerging subsequences were used to predict disordered regions in a protein sequence sliding window.

The predictor is based on emerging subsequences (ESs) which have high discriminating power, and it is more suitable to use ESs in classification analysis. Comparing with most existing disorder predictors which use a sliding window to map individual residues into a certain feature space, the ES-based predictor decreases the useless patterns for classification. The predictor using sliding window applies the feature selection for selecting more useful patterns. However, the ES-based predictor does not need to change the window size and prunes the generated patterns using feature selection methods.

The rest of the paper is organized as follows. Section 2 presents the method applied to the ES-based predictor using some examples. Section 3 shows the performance of the predictor and discusses the experiment results. Finally, we give some concluding remarks in Section 4.

2. Materials and methods

2.1. Emerging Subsequence

Sequence data are special data which have ordering properties. To discover the emerging pattern from sequence data, an emerging substring and a suffix tree-based framework for generating emerging substring were proposed by Sarah Chan in 2003 [36]. In this paper, to apply the emerging sequential patterns to protein sequence data, the emerging pattern was called an Emerging Subsequence (ES) and defined as being a part of a protein sequence that has a higher frequency of occurrence in the target class than in the source class. Emerging subsequences are more suitable for classifying protein sequences to the disordered sequences and ordered sequences than frequent sequential patterns which are often used in subsequence mining, because of the high discriminating power of emerging subsequences. Frequent sequential patterns only depend on the frequency of the subsequence in the target class.

However, the emerging subsequences do not only use the frequency of subsequence in the target class, but also compare with the background class.

2.2. Parameters

Two parameters: support (support count) and growthrate are used to generate the ESs. The support count is the number of subsequences in a target class, and the support is the rate of a subsequence among proteins that are included in a target class [given in Eq. (1)].

$$\text{support count}_k(s) = \text{The number of subsequences in class } k, \tag{1}$$

where k is a target class, disordered protein or ordered protein and s is a subsequence.

Growthrate of an ES is the ratio of support count or support which is contained in a target class to the support count or support which is contained in the background class [given in Eq. (2)].

$$\text{growthrate}\left(s\right) = \begin{cases} 0 & \text{if } \text{supp}_{\text{count 1}} = 0 \text{ and } \text{supp}_{\text{count 2}} = 0, \\ \infty & \text{if } \text{supp}_{\text{count 1}} = 0 \text{ and } \text{supp}_{\text{count 2}} > 0, \\ \dfrac{\text{supp}_{count2}}{\text{supp}_{count1}} & \text{otherwises.} \end{cases} \tag{2}$$

where $supp_{count1}$ is the value of support count of class 1, and $supp_{count2}$ is the value of support count of class 2.

However, a protein sequence usually contains more than one disordered region and a protein can also contain more than one subsequence. Consequently, an emerging subsequence also may be present several times in a protein sequence. The support count can be larger than the number of proteins in the target class and the support of a subsequence could be larger than "1". The support value cannot represent the rate in the total target class's protein sequences. Therefore, in this study, we applied the support count and growthrate as basis parameters unlikely in the study of basic emerging subsequence. In this case, support (support count) represents the frequency of a subsequence in target class, and growthrate represents the frequent standard of subsequence in target class which was compared to background class. In other words, an emerging subsequence of a class k is that a subsequence satisfies the threshold value of support count and growthrate value in class k.

In this work, the problem regarding prediction of disordered regions is to find the subsequences based on the support count and growthrate, which more frequently occur in the target class, and the disorder/order class, than in the background class, or order/disorder class, from the protein sequences whose structures need to be predicted.

2.3. Extraction of protein emerging subsequence

The protein emerging subsequence (Protein_ES) generator is a part of disordered region predictor that is used in this work. The Protein_ES generator is constructed based on a suffix tree which is used to arrange and search the sequence. The emerging subsequence mining algorithm was proposed by Sarah Chan and colleagues [36]. However, they proposed a single-class mining algorithm. In this work, we changed framework of the single-class mining algorithm to make it to be suitable for a two-class mining algorithm and the generation of disordered and ordered protein emerging subsequences.

A merged tree is a data structure which is based on a suffix tree [36] and can show all subsequence sequences, and also reveals the support count and growthrate value of each node in target class. In a merged tree, edges spell nonempty sequences and each node has at least two children apart from the leaf node. Every pathway from the root node to the leaf node is a suffix of sequences in a protein dataset. The purpose of constructing a merged tree is discovering all the subsequence of sequence in a sequence dataset, and it is also easy to calculate the support count and growthrate of all subsequences in the protein sequence dataset.

The process of generating a protein emerging sequence using a protein sequence dataset is as follows. At the beginning, there is just a root node in a merged tree. The root node does not represent anything. It is just used as a top node to connect all the nodes which do not have a parent node. A disordered sequence in a disordered sequence dataset is taken to compare to the root node's child nodes where every node represents an amino acid in the disordered protein sequence. If there is a node that is the same as compared to an amino acid in a disordered sequence, the disorder class's support counter of the node is added to one and compared to the next amino acid in the sequence with the pathway of the node's child nodes.

If there is no node that is the same as compared amino acids in the disordered sequence, the amino acids will create the new child node of root node, and the pathway of the nodes and new child node represents the different subsequence constructed by amino acids. The disorder class's support counter of the node added to one. Once the sequence is finished arranging, the next

sequence in the disordered sequence dataset starts being compared with the root node's child nodes of the merged tree. The ordered protein sequences dataset also use the same phases to upgrade the merged tree and to calculate the support count and growthrate value of emerging subsequence in the order class.

As we described, the amino acid is the unit of the protein sequence algorithm. Every node represents an amino acid, and the support counter value represents the frequency of the subsequence which is combined by amino acids in the pathway from the root node's child node to the appropriate node.

2.4. Identification of disordered region in proteins

The process is divided into two phases for identifying the disordered protein region in protein sequences using protein emerging subsequences. One is the phase of searching for disordered regions using disordered emerging subsequences (Disordered_ESs). The other is the pruning phase using ordered emerging subsequences (Ordered_ESs) to improve the prediction accuracy based on calculating contributions of protein emerging subsequences.

2.4.1. Searching disordered regions based on Disordered_ES

In this work, the method is used for discovering disordered regions by scanning the Disordered_ES using the sliding window technique in protein sequences and matching the right region as disordered regions. When amino acids in protein sequences were classified to disordered regions, the amino acids made a record of the support count and growthrate values of all Disordered_ESs which matched with the amino acids sequence.

2.4.2. Pruning disordered regions based on Ordered_ES and Score

The purpose of the pruning phase is to search the ordered regions that were incorrectly predicted as disordered regions by disordered emerging sequence. Consequently, the prediction accuracy is improved. In this phase, we also applied the sliding window to find the ordered region based Ordered_ES. The predicted disordered regions and ordered regions inevitably overlapped. In these cases, the parameter-score is applied to predict the disordered region. The score is proposed in the CAEP (classification by aggregating emerging patterns) [48] which applies the support and growthrate of emerging subsequence. It shows the sum of the contributions of the emerging subsequences in the target class (Eq. (3)). The formula is as follows.

$$\text{score}(a,k) = \sum_{a \subseteq s, s \in ES(k)} \frac{\text{growthrate}(s)}{\text{growthrate}(s)+1} * \text{supp\_count}_k(s) \quad (3)$$

where a is an amino acid which is contained in the overlapped region by Disordered_ES and Ordered_ES, s is an emerging subsequence, *ES(k)* is a dataset of emerging subsequences of class k. Namely, contribution is the value of as following formula,

$$\frac{\text{growthrate}(s)}{\text{growthrate}(s)+1} * \text{supp\_count}_k(s).$$

An example of the prediction method using the score of emerging subsequence is given in Table 1 and Figure 1. In the example, two overlap sequences "DSK" and "DD" exist in a protein sequence. For first overlap region "DSK", the scores are Score("D", D_ES) = 1.5 + 1.8 = 3.3 ("TTTLDSK" and "LDS") and Score("D", O_ES) = 1.6 ("DSKT"). Therefore "D" is a disordered region. For second amino acid "S", the scores are Score ("S", D_ES) = 1.5 + 1.8 = 3.3 ("TTTLDSK" and "LDS") and Score("S",O_ES) = 1.6 ("DSKT"). Therefore "S" is classified to the disordered region. Though the calculation, the classification result is that "TTTLDS" and "SKK" are the disordered regions.

3. Results

3.1. Dataset

In this study, the training data is extracted from DisProt (version 6.02, http://www.disprot.org/) and the Protein Data Bank (PDB). DisProt is a collection of disordered regions of proteins based on published literature descriptions. The PDB sequences were filtered using the culled PDB list to extract a high-quality and low-sequence identity subset. It has 694 proteins entries and 1539 disordered regions. Long disordered regions

Table 1. The example of disordered emerging subsequences and ordered emerging subsequences and their contribution values.

Disordered_Ess		Ordered_ESs	
Sequence	Contribution	Sequence	Contribution
TTTLDSK	1.5	DSKT	1.6
LDS	2.8	KT	2.8
DDSKK	1.5	TLDDD	1.3

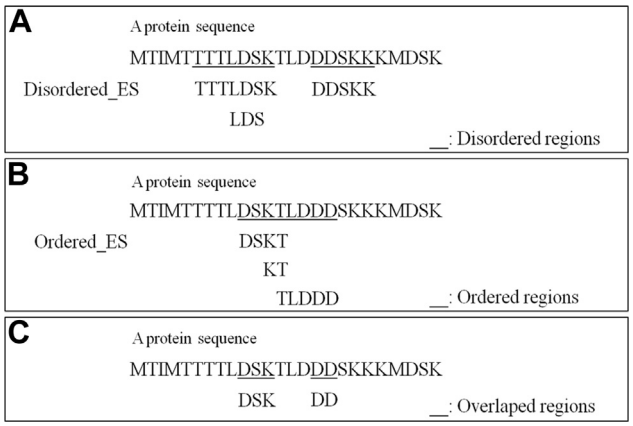

Figure 1. The example of prediction on overlap region in protein sequences. (A) Predicted disordered regions. (B) Predicted ordered region. (c) Overlapped regions of disordered emerging subsequences and ordered emerging subsequence.

(more than 30 amino acids) in DisProt 6.02 are used to train emerging subsequence based predictor to discover the long disordered emerging sequences, and it is denoted as the Long_Disorder dataset (LD). Short disordered regions (less than 30 amino acids) in DisProt 6.02 were used to train emerging subsequence based predictor to discover the short disordered emerging subsequences, and it is denoted the Short_Disorder dataset (SD). The ordered training data are extracted from PDB-Select-25, a representative set of protein data bank (PDB). This collection of ordered training set includes a total of 68,132 residues.

The CASP 9 and CASP 10 targets were used as an independent test dataset to blind test the performance of prediction. The CASP 9 dataset contains 108 sequences with a total of 21,230 residues, and the CASP 10 dataset contains 94 sequences with a total of 37,335 residues [49]. In this work, we randomly selected the 95 sequences in CASP 9 and CASP 10 which contain both disordered and ordered regions together for test data.

3.2. Protein emerging subsequences

The features used to predict disordered regions are: protein short disorder (SD) emerging subsequences and long disorder (LD) emerging subsequences generated by the protein emerging subsequence mining algorithm. When the ES is generated, two datasets, the background class dataset and the target class dataset, are needed. Therefore we set the dataset SD to the background class dataset, and set dataset LD to the target class dataset. Tables 2 and 3 show the sample of the disorder ESs generated. Here short_disorder/long_disorder ES means the ES is more frequent in short_disorder/long_disorder ES than in long_disorder/short_disorder ES. In Table 2, it is shown that the short_disordered region is different from long_disordered region as published [36].

3.3. Performance of ES_based disordered region predictor

To evaluate the performance of this approach, we compared with some existing predictors which have high accuracy.

3.3.1. The analysis and performance result

From Figure 2, we can determine that this approach is efficient for predicting the boundary of protein disordered regions. In the whole text dataset, it predicts 75% of boundary of disordered regions.

Table 2. Short_disorder ES.

Short_disorder ES	Support count	Growthrate
MEKVL	9	∞
FMEKV	9	∞
EKVL	9	4.5
KVLG	7	∞
AFMEK	7	∞
DPTI	6	∞
QEEY	6	6
YDPTI	6	∞
...	...	...

Table 3. Long_disorder ES.

Long_disorder ES	Support count	Growthrate
AKSPA	22	∞
EEEEG	15	∞
GQPHG	12	∞
WGQPH	12	∞
EEEED	11	11
GGWGQ	11	∞
DSDSD	11	∞
HGGGW	10	∞
...	...	...

Figure 2. Example of performance result.

Table 4. Performance accuracies of the protein disordered region on per-residue version (%).

	Sensitivity	Specificity	Precision	Accuracy
EMBL_hot	60.2	66.2	20.9	65.1
EMBL_coil	47.3	75.1	22.9	70.8
EMBL_remark	65.6	49.1	18.9	50.9
PONDR-VSL2	39.9	79.2	22.8	74.1
RONN	33.6	84.3	26.8	77.2
ES	44.7	85.3	30.7	79.4

3.3.2. The analysis of protein disordered region prediction by the per-residue method

For analysis on ES based predictor is more detail, the experiment on per-residue is proposed. Table 4 shows that the accuracies are generally high. In this experiment, we just compare with the predictors which applied the different properties of disordered regions, and obtain the highest accuracy [support = (2,2) growthrate = (2,3)]. Some papers reported that the sensitivity was the most important performance measure for disordered region predictor. If we change the parameters, the specificity of our proposed predictor also could reach 57.5% and the accuracy could reach 66% [support = (2,2), growthrate = (4,2)]. It means we can control the value of the parameters to obtain the necessary results.

Table 5. Performance accuracies of the protein disordered region on per-chain version (%).

Methods	Sensitivity	Specificity	Precision	Accuracy
EMBL_hot	40.8	79	20	74.6
EMBL_coil	64.4	47.4	13.6	49.3
EMBL_remark	21.1	93.3	28.7	85
PONDR-VSL2	37.6	82.8	21.9	77.6
RONN	34.1	84.3	26.8	77.1
ES	40.6	86.1	27.3	80.9

3.3.3. The analysis of the protein disordered region prediction by the per-chain method

When predicting the disordered region on per-chain version, the accuracies of all the predictors are not as good as per-chain version (Table 5). It is show that in the special case the predictor accuracy is much worse so that the overall accuracy is not higher than by using the per-chain method.

4. Disussion

Prediction of protein structures and functions, in particular identification of natively disordered and ordered regions of a protein, is always an important and challenging task. Although many predictors have been published, the study of protein disordered region prediction is ongoing because different prediction methods can find different disordered regions in a protein sequence. We have used a new approach to find the different disordered regions for more efficient and accurate prediction of protein structures. In this paper, we proposed a protein disordered region predictor was applied an emerging subsequence mining algorithm. An emerging subsequence, which has high discriminating power, is more suitable for classification analysis. The proposed prediction model uses a merged tree based on a suffix tree to discover the emerging subsequence from protein disordered and ordered sequence data, and

predicts the disordered region by identifying disorder emerging subsequence and ordered emerging subsequences using a sliding window in a protein sequence. Classification of the disordered regions and ordered region in a protein is according to the score of emerging subsequences. For testing the performance of the proposed predictor, we used the protein disordered sequence data from Disport 5.7 and ordered sequence data from the PDB. The extracted test data are from CASP 9 and CASP 10. The results show that this new approach guarantees high accuracy, and it is an efficient approach to predict the boundary of disordered regions compared with other methods.

The upgraded emerging subsequences-based predictor is appropriate to analyze protein structures and functions. We assumed that the emerging subsequences could discover regions of important biological significance and it could be used as part of meta-predictors. We also estimate that the disordered properties and emerging subsequences features could be used together to predict disordered regions and could obtain more meaningful features with high accuracy. Concurrently, it is also the topic of our future work. Regarding the number of disordered and ordered data, the parameters used in predictor are very different, and the way the parameters are set impacts the prediction accuracy. The discovery of the association between the parameters and the amounts of disordered and ordered data is for another future work.

Conflicts of interest

The authors declare no conflicts of interest.

Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (No.2013R1-A2A2A01068923), the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIP) (No. 2008-0062611), and Korea Biobank project (4851-307) of the Korea Centers for Disease Control and Prevention.

References

1. Uversky VN. Unusual biophysics of intrinsically disordered proteins. Biochim Biophys Acta 2013 May;1834(5):932–51.
2. Cozzetto D, Jones DT. The contribution of intrinsic disorder prediction to the elucidation of protein function. Curr Opin Struct Biol 2013 Jun;23(3):467–72.
3. Ekman D, Light S, Björklund Å, et al. What properties characterize the hub proteins of the protein-protein interaction network of *Saccharomyces cerevisiae*. Genome Biol; 2006. 1.7, 6, R45.
4. Apic G, Ignjatovic T, Boyer S, et al. Illuminating drug discovery with biological pathways. FEBS Lett 2005 Mar 21;579(8): 1872–7.
5. Gould CM1, Diella F, Via A, et al. ELM the status of the 2010 eukaryotic linear motif resource. Nucl Acids Res 2010 Jan; 38(Database issue):D167–180.
6. Oldfield CJ, Xue B, Van YY, et al. Utilization of protein intrinsic disorder knowledge in structural proteomics. Biochim Biophys Acta (BBA) - Proteins Proteomics 2013 Feb;1834(2):487–98.
7. Romero P, Obradovic Z, Kissinger CR, et al. Identifying disordered regions in proteins from amino acid sequences. IEEE Int Conf Neural Netw; 1997 Jun:90–5.
8. Huang T, He Z, Cui W, et al. A sequence-based approach for predicting protein disordered regions. Protein Peptide Lett 2013 Mar;20(3):243–8.
9. Ward JJ, Sodhi JS, McGuffin LJ, et al. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J Mol Biol 2004 Mar 26;337(3):635–45.
10. Peng K, Radivojac P, Vucetic S, et al. Length-dependent prediction of protein intrinsic disorder. BMC Bioinform 2006 Apr 17;7: 208.
11. Vullo A, Bortolami O, Pollastri G, et al. Spritz: a server for the prediction of intrinsically disordered regions in protein sequences using kernel machines. Nucleic Acids Res 2006 Jul 1;34(Web Server issue):W164–8.
12. Hirose S, Shimizu K, Satoru K, et al. POODLE-L: a two-level SVM prediction system for reliably predicting long disordered regions. Bioinformatics 2007 Aug 15;23(16):2046–53.
13. Ishida T, Kinoshita K. PrDOS: prediction of disordered protein regions from amino acid sequence. Nucleic Acids Res 2007 Jul; 35(Web Server issue):W460–4.
14. Mizianty MJ, Stach W, Chen K, et al. Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources. Bioinformatics 2010 Sep 15;26(18): i489–96. http://dx.doi.org/10.1093/bioinformatics/btq373.
15. Yang ZR, Thomson R, McNeil P, et al. RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. Bioinformatics 2005 Aug 15;21(16):3369–76.
16. Romero P, Obradovic Z, Kissinger CR, et al. Identifying disordered regions in proteins from amino acid sequences. IEEE Int Conf Neural Netw 1997 Jun;1:90–5.
17. Li X, Romero P, Rani M, et al. Predicting protein disorder for N-, C-, and internal regions. Genome Inform 1999;10:30–40.
18. Romero P, Obradovic Z, Li X, et al. Sequence complexity of disordered protein. Proteins 2001 Jan 1;42(1):38–48.
19. Liu J, Rost B. NORSp: predictions of long regions without regular secondary structure. Nucleic Acids Res 2003 Jul 1;31(13): 3833–5.
20. Peng K, Vucetic S, Radivojac P, et al. Optimizing long intrinsic disorder predictors with protein evolutionary information. J Bioinform Comput Biol 2005 Feb;3(1):35–60.
21. Linding R, Jensen LJ, Diella F, et al. Protein disorder prediction: implications for structural proteomics. Structure 2003 Nov;11(11): 1453–9.
22. Ward JJ, McGuffin LJ, Bryson K, et al. The DISOPRED server for the prediction of protein disorder. Bioinformatics 2004 Sep 1; 20(13):2138–9.
23. Cheng J, Sweredoski MJ, Baldi P. Accurate prediction of protein disordered regions by mining protein structure data. Data Mining Knowl Disc 2005 Nov;11:213–22.
24. Vucetic S, Brown CJ, Dunker AK, et al. Flavors of protein disorder. Proteins 2003 Sep 1;52(4):573–84.
25. Obradovic Z, Peng K, Vucetic S, et al. Predicting intrinsic disorder from amino acid sequence. Proteins 2003;53(Suppl 6):566–72.
26. Obradovic Z, Peng K, Vucetic S, et al. Exploiting heterogeneous sequence properties improves prediction of protein disorder. Proteins 2005;61(Suppl. 7):176–82.

27. Uversky VN, Gillespie JR, Fink AL. Why are "natively unfolded" proteins unstructured under physiologic conditions. Proteins 2000 Nov 15;41(3):415−27.
28. Prilusky J, Felder CE, Zeev-Ben-Mordehai T, et al. FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. Bioinformatics 2005 Aug 15;21(16):3435−8.
29. Han P, Zhang X, Feng ZP. Predicting disordered regions in proteins using the profiles using amino acid indices. BMC Bioinformatics 2009;10(Suppl. 1):S42.
30. Bulashevska A, Eils R. Using Bayesian multinomial classifier to predict whether a given protein sequence is intrinsically disordered. J Theor Biol 2008 Oct 21;254(4):799−803.
31. Linding R, Jensen LJ, Diella F, et al. Protein disorder prediction: implications for structural proteomics. Structure 2003 Nov;11(11): 1453−9.
32. Peng K, Vucetic S, Radivojac P, et al. Optimizing long intrinsic disorder predictors with protein evolutionary information. J Bioinform Comput Biol 2005 Feb;3(1):35−60.
33. Peng K, Radivojac P, Vucetic S, et al. Length-dependent prediction of protein intrinsic disorder. BMC Bioinformatics 2006 Apr 17;7:208.
34. Jones DT, Ward JJ. Prediction of disordered regions in proteins from position specific score matrices. Proteins 2003;53(Suppl 6): 573−8.
35. Schlessinger A, Liu J, Rost B. Natively unstructured loops differ from other loops. PLoS Comput Biol 2007;3:e140.
36. Chan S, Kao B, Yip CL, et al. Mining emerging substrings. Conf. on Database Systems for Advanced Applications; 2003. p. 119−26.
37. Dosztányi Z, Csizmok V, Tompa p, et al. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics 2005 Aug 15; 21(16):3433−4.
38. Dosztanyi Z, Csizmok V, Tompa P, et al. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. J Mol Biol 2005 Apr 8;347(4):827−39.
39. Garbuzynskiy SO, Lobanov MY, Galzitskaya OV. To be folded or to be unfolded? Protein 2004 Nov;13(11):2871−7.
40. McGuffin LJ. Intrinsic disorder prediction from the analysis of multiple protein fold recognition models. Bioinformatics 2008 Aug 15;24(16):1798−804.
41. McGuffin LJ. The ModFOLD server for the quality assessment of protein structural models. Bioinformatics 2008 Feb 15;24(4): 586−7.
42. Schlessinger A, Punta M, Yachdav G, et al. Improved disorder prediction by combination of orthogonal approaches. PLoS ONE 2009 Feb;4:e4433.
43. Ishida T, Kinoshita K. Prediction of disordered regions in proteins based on the meta approach. Bioinformatics 2008;24:1344−8.
44. Xue B, Dunbrack RL, Williams RW, et al. PONDR-FIT: A meta-predictor of intrinsically disordered amino acids. Biochim Biophys Acta 2008 Jun 1;24(11):1344−8.
45. Fan X, Kurgan L. Accurate prediction of disorder in protein chains with a comprehensive and empirically designed consensus. J Biomol Structure Dyn 2014;32(3):448−64.
46. Mizianty MJ, Uversky V, Kurgan L. Prediction of intrinsic disorder in proteins using MFDp2. Protein Struct Prediction Methods Mol Biol 2014;1137:147−62.
47. Moran O, Roessle MW, Mariuzza RA, et al. Structural features of the full length adaptor protein GADS in solution determined using small-angle X-ray scattering. Biophys J 2008 Mar 1;94(5): 1766−72.
48. Dong G, Zhang X, Wong L, et al. CAEP: classification by aggregating emerging patterns. Lecture Notes Comp Sci 1999;1721.
49. Monastyrskyy B, Kryshtafovych A, Moult J, et al. Assessment of protein disorder region predictions in CASP10. Proteins. Struct Function Bioinform 2014 Feb;82(Suppl 2):127−37.