Briefings in Bioinformatics, 2017, 1–17

doi: 10.1093/bib/bbx126
Paper

A comprehensive review and comparison of existing computational methods for intrinsically disordered protein and region prediction

Yumeng Liu, Xiaolong Wang and Bin Liu*

*Corresponding author: Bin Liu, Harbin Institute of Technology Shenzhen Graduate School, HIT Campus Shenzhen University Town, Xili, Shenzhen, 518055, China. Tel.: +86 0755-86011630; E-mail: bliu@hit.edu.cn

Abstract

Intrinsically disordered proteins and regions are widely distributed in proteins, which are associated with many biological processes and diseases. Accurate prediction of intrinsically disordered proteins and regions is critical for both basic research (such as protein structure and function prediction) and practical applications (such as drug development). During the past decades, many computational approaches have been proposed, which have greatly facilitated the development of this important field. Therefore, a comprehensive and updated review is highly required. In this regard, we give a review on the computational methods for intrinsically disordered protein and region prediction, especially focusing on the recent development in this field. These computational approaches are divided into four categories based on their methodologies, including physicochemical-based method, machine-learning-based method, template-based method and meta method. Furthermore, their advantages and disadvantages are also discussed. The performance of 40 state-of-the-art predictors is directly compared on the target proteins in the task of disordered region prediction in the 10th Critical Assessment of protein Structure Prediction. A more comprehensive performance comparison of 45 different predictors is conducted based on seven widely used benchmark data sets. Finally, some open problems and perspectives are discussed.

Key words: intrinsically disordered proteins and regions; physicochemical-based method; machine-learning-based method; template-based method; meta method

Introduction

Proteins and protein regions without a stable three-dimensional structure are called intrinsically disordered proteins (IDPs) and intrinsically disordered regions (IDRs) [1, 2]. In fact, there are many terms used to describe these proteins or regions, such as intrinsically disordered [3], intrinsically unstructured [4], natively unfolded [5] and natively disordered [6]. In this article, we use IDP and IDR. Figure 1 shows a schematic view of a protein 1QME in Protein Data Bank (PDB) with four IDRs (i.e. missing electron densities in X-ray structure).

IDPs/IDRs are widely distributed in known natural proteins, particularly in eukaryotic proteins [8–10], and many significant biological functions have also been confirmed to be associated with them. These functions may come from different IDP/IDR states or transitions among these states [11–13], which include disordered state, molecular recognition and binding to partner molecules. Some examples of these functions are flexible linker, assembler, cellular signal transduction, protein phosphorylation [11, 14]. Besides, IDPs/IDRs are correlated with a broad range of human diseases, such as cancer [15], genetic diseases [16], cardiovascular disease [17], amyloidoses, neurodegenerative

Yumeng Liu is a PhD candidate at the School of Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate School, China. Her research areas include bioinformatics and machine learning.
Xiaolong Wang, PhD, is a professor at the School of Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate School, China. His research areas include nature language processing, bioinformatics and artificial intelligence.
Bin Liu, PhD, is a professor at the School of Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate School, China. His expertise is in bioinformatics, nature language processing and machine learning.
Submitted: 19 July 2017; **Received (in revised form):** 5 September 2017

diseases [18], synucleinopathies and Alzheimer's disease [19, 20]. Therefore, the identification of IDPs/IDRs is important, and will contribute to the drug design [21, 22].

Many experimental techniques have been used to identify IDPs/IDRs [23, 24], for example, nuclear magnetic resonance (NMR), X-ray crystallography, circular dichroism (CD) spectroscopy, small-angle X-ray scattering (SAXS) and single-molecule fluorescence resonance energy transfer (smFRET). However, these methods are time-consuming and costly. For the prevalence of IDPs/IDRs, a series of computational predictors have been presented [1, 2, 24–27]. All these predictors capture the characteristics of IDPs/IDRs to improve the predictive performance. One of the most commonly used characteristics is the length of IDRs. The IDRs can be divided into long disordered region (LDR) and short disordered region (SDR) [28–31] based on their lengths. Usually, the LDRs have >30 amino acid residues, and SDRs have ≤30 amino acid residues. The sequence position of IDRs in proteins is another important feature because previous studies observed that the distributions of IDRs are different in the N-terminal, C-terminal and internal regions, and some predictors are built for predicting these regions [32, 33]. Besides, there are also some predictors that consider the definition of disorder when building the prediction model [34, 35].

There are several review papers [1, 2, 25–27, 36–39] published during the past decade, which have comprehensively reviewed the developments of the IDP/IDR predictors. All these papers have played a role in simulating the development of this important field, especially two recent review papers [38, 39] have made great contributions, for example, the review paper [39] not only reviews the updated IDP/IDR predictors, but also discusses the molecular functions of disordered proteins and their predictors. Inspired by these important review papers, in this study, we conduct a comprehensive review of the computational predictors in this field, especially focusing on the recently proposed predictors. To fairly evaluate the performance of different predictors, 40 state-of-the-art methods are tested on the target proteins in the task of disordered region prediction in CASP10 for directly performance comparisons. For the unreported methods, their results are obtained by running their Web servers or stand-alone programs. Furthermore, more comprehensive performance comparisons are conducted, and 45 different predictors are compared based on seven widely used benchmark data sets [40–45]. To our best knowledge, the performance evaluation reported in this article is the most comprehensive and updated performance comparison.

Databases and benchmark data sets of IDPs/IDRs

In the past decades, owing to the prevalence and importance of IDPs/IDRs, several databases of IDPs/IDRs have been constructed, including DisProt [11], IDEAL [46], MobiDB [47, 48] and D^2P^2 [49].

DisProt [11] is one of the most commonly used databases for IDP/IDR prediction, which manually collects the IDPs/IDRs with experimental evidence reported. These IDPs/IDRs are characterized by various experimental techniques. Among these techniques, X-ray crystallography and multidimensional NMR are considered as the 'primary techniques' for providing evidence of IDPs/IDRs, and several alternative biochemical and biophysical approaches are treated as 'secondary techniques' for providing orthogonal information of IDPs/IDRs. For the current version DisProt 7.0, there are 803 entries and 2167 IDRs with a total number of 92 432 amino acids with experimental and functional annotations. In these IDRs, there are >800 SDRs and around 1200 LDRs.

In addition, IDEAL [46] and MobiDB [47, 48] databases are two useful databases for IDP/IDR prediction. The IDPs/IDRs in IDEAL are also experimentally verified. MobiDB collects IDRs data from three levels [47]: The first level is manually curated data extracted from the DisProt [11]; the second level is indirect data inferred from PDB-NMR and PDB-xray structures; the last level is predicted data by using 10 predictors, making MobiDB able to annotate disorder annotations for any protein without reliable data. For each data source (i.e. DisProt, PDB and predictors), three possible states, including structure, disorder and ambiguous, are assigned for each residue. Then, a consensus annotation is made by combining all the data sources. Moreover, MobiDB also provides a consensus annotation for LDRs. Another database D^2P^2 contains IDPs/IDRs predicted by nine predictors based on proteins from 1765 complete proteomes [49]. A summary of IDP/IDR databases is shown in Table 1.

Besides, the IDPs/IDRs acquired from PDB are also widely used by many predictors [25]. In PDB database, those regions solved by X-ray crystallography and missing electron densities are defined as IDRs [7], otherwise, they are ordered regions.

Methods

As the first predictor proposed by Romero *et al.* in 1997 [55], during the past two decades, many computational predictors have been proposed for identification of IDPs/IDRs. According to the different strategies, these methods can be divided into four categories, including physicochemical-based method, machine-learning-based method, template-based method and meta method. The strategies of these four categories are summarized as follows. Physicochemical-based method is based on physical principles of protein sequences, and machine-learning-based method constructs classification models or sequence labeling models based on machine learning algorithms, template-based method builds predictor by searching known structure homologous proteins and meta method integrates the outputs of different predictors. However, please note that these four categories are not strictly different. In fact, predictors in one category may also use the techniques from other categories, for example, some physicochemical properties used by physicochemical-based methods are used as features to construct the machine-learning-based methods or meta methods.

Physicochemical-based methods

The physicochemical-based method predicts IDPs/IDRs based on the physical principles, which affect protein folding directly on the sequence. These physical principles include the propensities of specific amino acid composition, the combination of net charge and hydrophobicity, the interresidue contacts, etc. For example, Uversky *et al.* utilize the combination of low hydrophobicity and high net charge to distinguish whether a protein is an IDP [56]. Later, based on this principle, FoldIndex [57] is designed to identify IDRs through pre-defined sliding window. GlobPlot [51] uses a parameter P to predict IDRs, and its basic hypothesis is that the tendency for disorder can be characterized as $P = RC\text{-}SS$, where RC and SS represent the propensity for a given amino acid to be in 'random coil' and regular 'secondary structure', respectively. The basic algorithm behind GlobPlot is a sum function of P, which is simple and fast. Other methods

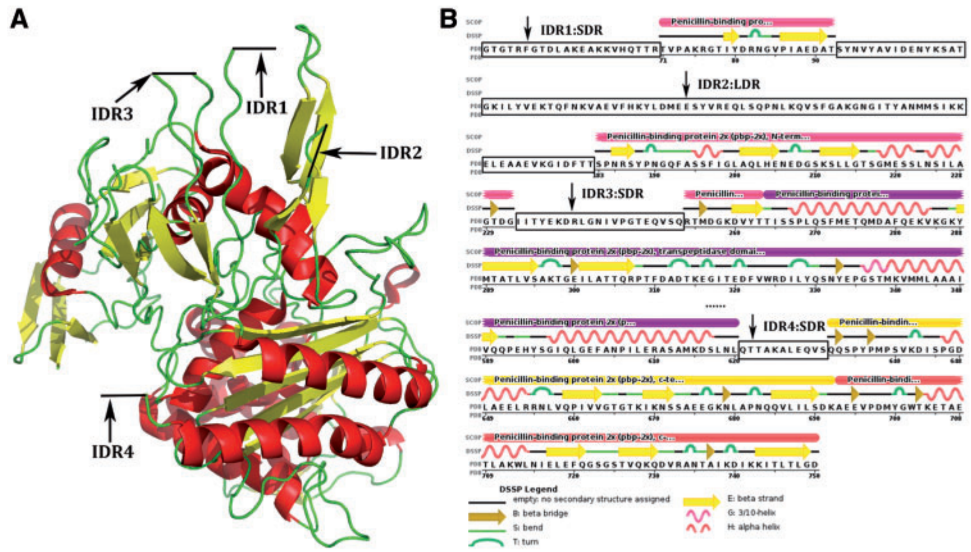

Figure 1. A schematic view of a protein 1QME in PDB with four disordered regions. Subfigure (**A**) shows the three-dimensional structure of this protein. The IDRs are shown as straight lines. Subfigure (**B**) shows its corresponding chain view, where the blank space represents the IDRs. This subfigure is from PDB Web site [7]. A colour version of this figure is available at BIB online: https://academic.oup.com/bib.

Table 1. The summary of IDP/IDR databases

Name	Version	**Method**[a]	Description	Web sites
DisProt [11]	v 7.03 26 September 2016	Experiment	Contains 803 entries and 2167 IDRs, with 92 432 amino acids	http://www.disprot.org/
IDEAL [46]	17 March 2017	Experiment	Contains 838 entries and 8501 non-redundant IDRs	http://www.ideal.force.cs.is.nagoya-u.ac.jp/IDEAL/
MobiDB [47][b]	v 2.3 July 2014	Experiment prediction	Combines experimental and predicted data into a consensus annotation. It contains disorder annotations for 80 370 243 entries	http://mobidb.bio.unipd.it/
D^2P^2 [49][c]	2012	Prediction	provides IDP/IDR predictions made by 9 different predictors across 1765 complete genomes containing 10 429 761 sequences from 1256 distinct species	http://d2p2.pro/

[a] Methods for identifying the IDPs/IDRs. Experiment means the IDPs/IDRs are identified by experimental techniques, for example X-ray crystallography and multidimensional NMR. Prediction means that the IDPs/IDRs are predicted by computational methods.

[b] The results are identified by 10 predictors, including two variants of IUPred [50], Globplot [51], two variants of DisEMBL [34], three variants of Espritz [35], RONN [52] and VSL2b [29].

[c] The results are predicted by 9 predictors, including PONDR VL-XT [53], PONDR VSL2 [29], PrDOS [54], PV2 [127], three variants of Espritz [35] and two variants of IUPred [50].

are based on interresidue contacts, taking structure and energy into account [50, 58–60]. For example, IUPred [50, 60] assumes that IDPs/IDRs do not have the capacity to form sufficient interresidue interactions to overcome the entropy loss during folding. The prediction of IUPred is carried out by estimating the capacity of polypeptides to form stabilizing contacts. The estimated energy for each residue not only considers its chemical type, but also considers its potential interaction with partners. The parameters of IUPred are derived from a global protein data set wihtout using IDPs [60]. Similar to IUPred, FoldUnfold [5, 58] designs a new parameter, called mean packing density, to represent the average contact number of residues within a given distance in a protein. It has been proved that regions with weak expected packing density correspond to the IDRs.

Physicochemical-based methods are efficient with low computational cost. Furthermore, their predicted results are easy to be interpreted. Therefore, additional useful information can be provided to the researchers who are interested in studying the IDPs/IDRs. The physicochemical properties can also be used as features for the machine-learning-based methods and meta methods. However, only based on one feature prevents their performance improvement.

Machine-learning-based methods

To overcome the disadvantages of physicochemical-based methods, methods based on machine learning techniques have been proposed. Compared with physicochemical-based method, these methods effectively use positive and negative samples to distinguish order and disorder, and incorporate various features. As an important component in machine learning methods, feature extraction techniques can be mainly divided into the three categories in IDP/IDP prediction. The first category is the sequence properties, such as the amino acid composition propensity [29, 53], the amino acid flexibility [58] and the low complexity [29]. The second category is the evolutionary information obtained from sequence profile [29, 30, 35, 61]. The last category is the structural information, such as secondary structure information [23, 61, 62], solvent accessibility [30, 61, 62], torsion angle fluctuation [30, 62]. Besides, various features considering the sequence order effects are also used for constructing the predictors [63–68]. The machine-learning-based methods can be further divided into two groups based on different machine learning models, including classification model and sequence labeling model.

Classification models

Traditional machine learning classification models can only handle fixed-length feature vectors. Classification models are trained in a supervised manner by using both the positive and negative samples, and then predict the unseen samples based on the trained model. The identification of IDRs is a binary classification problem, aiming to distinguish whether an amino acid residue is ordered or disordered one. The key of these methods is how to convert the proteins into fixed-length feature vectors. However, it is never an easy task because proteins have different lengths. In this regard, the sliding window technique is used to incorporate the context information of target residue by considering its neighboring residues. Some well-known classification algorithms have been applied to construct the predictors, such as Support Vector Machine (SVM), Neural Network (NN), Random Forest, Radial Basis Function Network (RBFN) and Logistic Regression (LR). The flowchart of machine-learning-based method is shown in Figure 2.

PONDR is the first predictor based on classification model [55], which constructs predictors based on NNs for different types of IDR, including SDR (7–21 residues), MDR (22–44 residues), LDR (45 or more residues) and all lengths of IDRs. This study shows that IDRs with different lengths have different characteristics, and predictors trained with one type of IDR cannot accurately predict other types of IDRs. Therefore, the following studies design predictors only for predicting one type of IDR, such as LDR or SDR, and some predictors are proposed, such as PONDR VL-XT [53], PONDR VL3, PONDR VL3E [69], POODLE-L [70], POODLE-S [32], DRaai-L [71], DRaai-S [71], Spritz [31] and SLIDER [72]. VL-XT [33, 53] is designed for predicting LDRs, containing three NN predictors optimized for N-terminal, C-terminal and internal regions, respectively. In VL-XT, composition-based and property-based attributes are selected as features, including flexibility, hydropathy, coordination number, net charge and specific amino acid composition. The VL3 [69] model is an ensemble of NNs, each of which is trained with a balanced data set randomly sampled from the training set. As VL-XT shows that low-complexity regions are more likely to be IDRs, VL3 adds the local sequence complexity as input combined with amino acid frequencies and flexibility. The VL3E predictor is a combination of VL3H and VL3P [69]. Different from VL3, VL3H searches for homologues proteins of IDPs to increase the number of training set, and VL3P adds profile feature as input. POODLE-L [70] is a complex predictor designed for predicting LDRs, which integrates 10 two-level SVM models, and uses the physicochemical properties of amino acids as features. For POODLE-L, SVM predicts the disorder probability of sequence segment at the first level, and then the second-level SVM calculates the disorder probability of each residue. In contrast to POODLE-L, POODLE-S [32] is designed for predicting SDRs. Because the propensities of amino acid composition information are different at the N-terminal, C-terminal, and internal regions of proteins, POODLE-S trains seven predictors for different regions along the protein sequences based on the Position-Specific Substitution Matrix (PSSM) and physicochemical properties. Spritz [31] contains two SVM-based predictors, one for SDRs and another for LDRs, which improves the predictive performance by using the different features of LDRs and SDRs, including amino acid frequencies, and the secondary structures predicted by the Porter server [73]. LDRs play an important role in various cellular functions and target selection. Owing to the lack of methods that directly predict proteins with LDRs, SLIDER [72] is proposed, which uses a carefully chosen set of features to construct a LR model for predicting whether a given protein contains LDRs. SLIDER is an accurate and fast predictor, which is suitable for high-throughput, genome-scale applications.

Because the aforementioned predictors are designed for predicting only one type of IDR (LDRs or SDRs), they cannot work well for predicting both LDRs and SDRs. Therefore, some predictors have been proposed to predict both LDRs and SDRs, which can be divided into two classes. Methods in the first class combine the region-specific predictors so as to predict both LDRs and SDRs, such as VSL1 [28], VSL2 [29] and SPINE-D [30]. VSL1 [28] is a two-layer model with three LR models. The first layer contains two specialized predictors, one is VSL1-L optimized for predicting LDRs and another is VSL1-S optimized for predicting SDRs. At the second layer, a meta-predictor VSL1-M is constructed to assign weights for combining the two specialized predictors in the first layer. VSL1 greatly improves the predictive performance for both SDRs and LDRs. To further explore the influence of length-dependent amino acid compositions and sequence properties of IDRs on the predictive performance, VSL2-M1 and VSL2-M2 are proposed [29]. VSL2-M1 has the same architecture as that of VSL1, but its component predictors are constructed by using SVMs instead of LRs. Compared with VSL-M1, VSL-M2 at the second layer uses a different method to merge two specialized predictors. SPINE-D [30] is based on NN with two-hidden-layer neural network, and an additional one-layer filter for smoothing prediction results. It first makes prediction for each residue with three possible states instead of two states, including ordered residue, residue in SDRs and residue in LDRs, and then converts the three states into two states by simply adding the probabilities in SDRs and LDRs. The input of SPINE-D includes residue-level features (physical parameters, PSSM, predicted secondary structure and predicted solvent accessibility) and window-level features (amino acid composition, local compositional complexity and predicted secondary structure content). Methods in the second class are trained with all IDRs, ignoring their differences, such as DISOPRED [74, 75], DISOPRED2 [8], DisEMBL [34], DisPSSMP [76], DisPSSMP2 [77], Dispredict [62] and svmPRAT [78]. Inspired by the secondary structure prediction model PSIPRED [79], DISOPRED [74, 75] is proposed to explore the role of sequence profile in the IDP/IDR prediction, which uses the sequence profile generated by PSI-BLAST [80] as the input of NN. Later, DISOPRED2 [8] improves the performance of DISOPRED by using SVMs instead of NNs. DisEMBL [34] is designed based on three NN models trained with three data sets with different definitions of disorder. DisPSSMP [76] combined RBFN and a matrix PSSMP, which is a condensed PSSM according to different physicochemical properties. Later, to further improve its predictive performance, a two-stage predictor DisPSSMP2 [77] is proposed, whose first stage adopts the same model architecture as that in DisPSSMP, and the second stage collects the outputs of the first stage by using sliding windows to

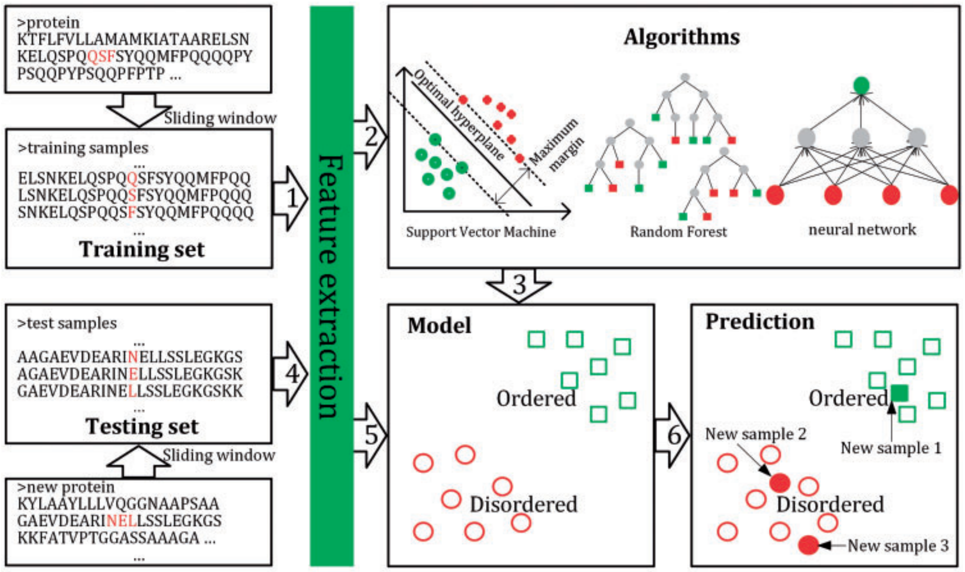

Figure 2. The flowchart of IDP/IDR prediction based on classification models. The target amino acids are shown in red. Training samples are mapped into a feature space by using feature extraction techniques, and then fed into classifiers to train prediction models for predicting unseen samples. A colour version of this figure is available at BIB online: https://academic.oup.com/bib.

generate the final prediction. Another predictor Dispredict [62] is based on SVMs with Radial Basis Function kernel. Dispredict uses three kinds of features, including sequence information, evolutionary information and the structural information.

Sequence labeling models

Sequence labeling models are trained in a supervised manner by using both the positive and negative samples, whose input is an observation sequence and the output is a labeling sequence. Their input is protein sequence, and the output is the labeling sequence, where each residue is labeled as ordered or disordered residue. Some well-known sequence labeling models have been applied to this field, such as Recurrent Neural Network (RNN), Conditional Random Field (CRF), Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM). DISpro [61] is based on RNN, and its input is a one-dimensional array *I*, whose dimension is equal to the length of protein sequence encoded by profile, secondary structure and relative solvent accessibility. The output of DISpro is a one-dimension array O and each dimension O_i represents the disorder probability of residue at position *i*. DISpro can handle variable length of proteins, and can capture the contextual information along proteins. Therefore, the output O_i is decided by long-ranged contextual information, rather than a local fixed-length window information centered at position *i*. Later, DISpro is improved by using flexible thresholds, leading to better tradeoff between specificity and sensitivity [81]. OnD-CRF [82] is based on CRF, whose features are amino acid sequence information and predicted secondary structure information. Compared with SVM-based and FNN-based models, OnD-CRF is able to incorporate the interrelation information among neighboring residues. ESpritz [35] is an ensemble of three NNs for predicting the N-terminal, internal and C-terminal of proteins, respectively. Recently, deep learning techniques have been widely applied to various tasks in bioinformatics and achieve high performance [83–85]. Some predictors for predicting IDPs/IDRs are constructed based on deep learning approaches. For example, DeepCNF-D [86] is a combination of Deep Convolutional Neural Network (DCNN) and CRF, which not only considers the relationships between sequence and structure in a hierarchical manner, but also incorporates the correlation among adjacent residues. Later, AUCpreD [87] with the same model as DeepCNF-D improves the performance of DeepCNF-D by maximizing AUC instead of maximum-likelihood in the process of training the model. SPOT-disorder [43] is a deep bidirectional LSTM RNN model with three-hidden-layer BRNN, including a recurrent feed-forward layer and succeeded by two LSTM cell layers. The results of SPOT-disorder show that it can achieve good performance for predicting both SDRs and LDRs. The reason is that LSTM can capture non-local, long-range interactions, and the bidirectional network is able to capture the forward and backward information along the protein sequences.

Compared with the classification models, the sequence labeling models are able to incorporate the interrelation information along the whole protein sequences. Combined with deep learning techniques, their performance is further improved. The reason is that RNN and LSTM can automatically recognize contextual information of the target residue, and capture both the local and global contextual information of proteins. A comparison between classification models and sequence labeling models is shown in Figure 3.

Template-based methods

The template-based method predicts IDRs by using homologous known structure proteins. These methods first try to search for homologues proteins with known structures (i.e. templates) [88], and then analyze the search results and predict the IDRs. For example, in PrDOS [54] and GSmetaDisorder3D [89], template-based predictors are used as component predictors rather than independent predictors because of their different design principles.

The advantage of template-based methods is that the predicted results are easy to interpret. However, the matching homologous proteins of target protein have different reliabilities, and in some cases, homologous sequences cannot be

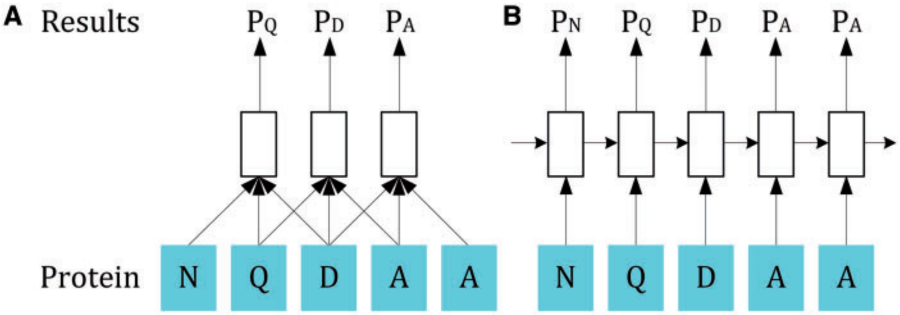

Figure 3. The comparison between classification model and sequence labeling model. Subfigure (**A**) shows that the prediction of classification model using local sequence information extracted from sliding windows. Subfigure (**B**) shows that the sequence labeling model is able to capture the interaction among all the residues in a protein.

detected. The flowchart of template-based method is shown in Figure 4.

Meta methods

As discussed above, for different computational predictors, they have their own advantages and disadvantages. Therefore, some meta methods have been proposed to combine various predictors into one model to further improve the predictive performance [1]. According to different fusion strategies, they can be divided into two categories, including linear fusion and machine-learning-based fusion. The flowchart of the meta methods is shown in Figure 5.

Linear fusion

Linear fusion combines the results of different methods by a weighted voting strategy, such as PrDOS [54], MULTICOM [90], CSpritz [91], MetaDisorder [89] and MobiDB-lite [92]. PrDOS [54] has two predictors, one is a SVM-based model trained with PSSMs and the other is a template-based predictor. The final prediction is based on averaging the weighted results of the two predictors. Similar to PrDOS, MULTICOM [90] removes the weak predictors with low performance in the CASP8, and then combines the remaining predictors. MetaDisorder [89] contains three meta-predictors, including FloatCons, GSmetaDisorderMD and GSmetaDisorderMD2. FloatCons adopts 13 predictors, and the final consensus prediction is made by summing the weighted prediction scores of all predictors. GSmetaDisorderMD is constructed by combining FloatCons and a template-based predictor GSmetaDisorder3D, and the genetic algorithm is used to optimize the weights of the predictors. GSmetaDisorderMD2 is a variant of GSmetaDisorderMD, in which a different scoring function is adopted during genetic algorithm optimization. Recently, MobiDB-lite [92] is constructed based on eight different predictors, whose final consensus prediction is determined by voting.

Machine-learning-based fusion

Compared with linear fusion, machine-learning-based fusion uses the prediction results of individual predictors as features to construct predictors based on machine learning methods, such as metaPrDOS [93], MD [94], PONDR FIT [3], MFDp [95], MFDp2 [96], DisCop [44] and DISOPRED3 [97]. MetaPrDOS [93] uses the results from seven predictors to construct a meta-predictor with two steps. The first step is to collect prediction results from seven predictors. The second step is to integrate the collected results, and determine the disorder tendency for each residue based on SVM. MD [94] is a meta-predictor based on NN model requiring two inputs. The first input is the prediction results made by four predictors of IDP/IDR. The second input is sequence properties. MD achieves the best AUC (0.821), the second highest ACC (0.743) and the second highest MCC (0.444) in the evaluation [40]. Another meta-predictor PONDR FIT [3] is constructed based on six predictors. For each target residue, the prediction results of six predictors are fed into a NN with 20 hidden units for prediction. MFDp [95] is an ensemble of three SVM models for predicting SDRs, LDRs and generic IDRs, respectively. These three models use different parameters and features, including the results of three complementary disorder predictors, and some other information derived from sequence. The results predicted by the SVM-based predictor with the highest probability among the three predictors are considered as the final prediction results. MFDp obtains the best ACC (0.757), MCC (0.451) and AUC (0.821) in the evaluation [40]. It also achieves the highest ACC (0.795), the second highest MCC (0.553) and the second highest AUC (0.876) in another evaluation [44], and also shows relative better performance in other evaluations [42, 43, 98]. Later, MFDp2 [96] improves the performance of MFDp by adopting two additional predictors, one is predictor of disordered content DisCon [99] and another is an alignment-based predictor. The main idea of MFDp2 is to adjust the combination prediction results made by MFDp and alignment-based predictor to match the content predicted by DisCon. Another method DisCop [44] uses a rational design to construct a meta-predictor, which selects the best performance set of predictors from 20 basic predictors. The final selected methods include Espritz (the DisProt and X-ray versions) [35], Cspritz (the long version) [91], SPINE-D [30], DISOPRED2 [8], MD [94] and DISOclust [100]. The prediction results of these methods are then combined by using a regression model. DisCop shows the best MCC (0.571) and the best AUC (0.882) in the evaluation [44]. DISOPRED3 [97] is a two-layer model based on DISOPRED2. The first layer contains three predictors including DISOPRED2, a specialized predictor of LDRs and a nearest neighbor predictor, and the second layer is a NN model integrating the first layer prediction results.

By combining different methods, meta-predictors achieve the state-of-the-art performance. However, their computational cost is high, preventing their applications to data sets with high number of proteins.

Webservers and stand-alone tools

Because of the importance of IDP/IDR prediction, some web servers or stand-alone tools have been established, which are listed in Table 2. By using these severs or tools, the users can easily reproduce their reported results and make new

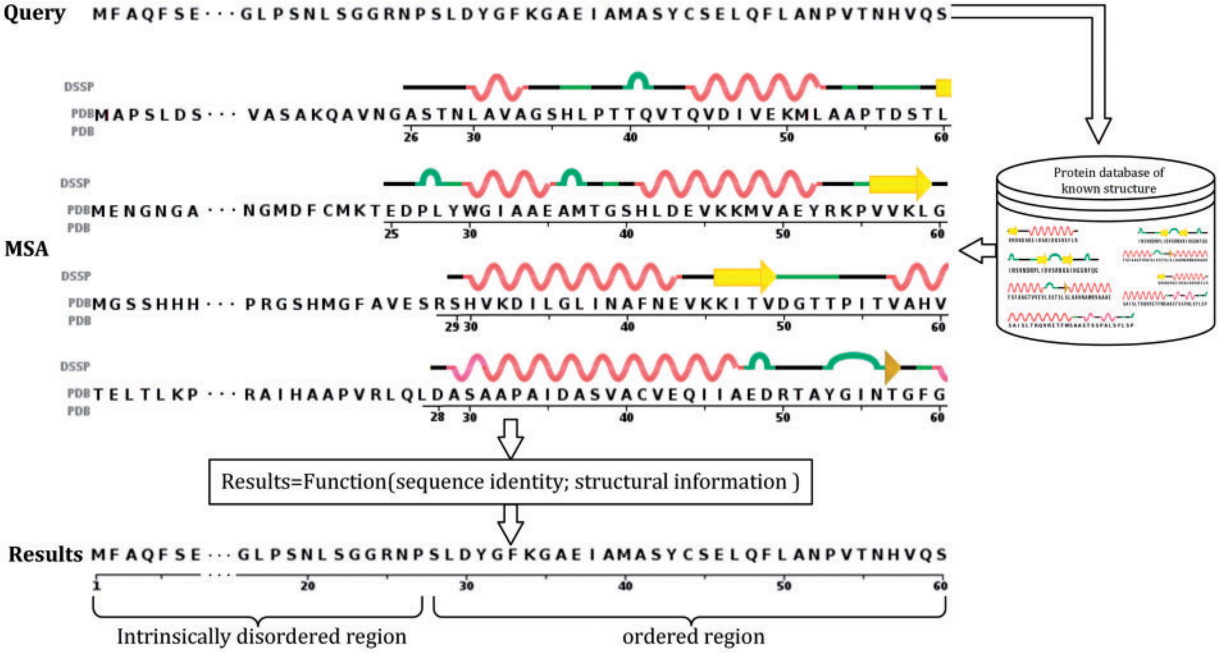
Query MFAQFSE···GLPSNLSGGRNPSLDYGFKGAEIAMASYCSELQFLANPVTNHVQS
DSSP
PDB MAPSLDS··· VASAKQAVNGASTNLAVAGSHLPTTQVTQVDIVEKMLAAPTDSTL
PDB
26 30 40 50 60
DSSP
PDB MENGNGA···NGMDFCMKTEDPLYWGIAAEAMTGSHLDEVKKMVAEYRKPVVKLG
PDB
25 30 40 50 60
MSA
DSSP
PDB MGSSHHH··· PRGSHMGFAVESRSHVKDILGLINAFNEVKKITVDGTTPITVAHV
PDB
29 30 40 50 60
DSSP
PDB TELTLKP···RAIHAAPVRLQLDASAAPAIDASVACVEQIIAEDRTAYGINTGFG
PDB
28 30 40 50 60
Protein database of known structure
Results=Function(sequence identity; structural information)
Results MFAQFSE···GLPSNLSGGRNPSLDYGFKGAEIAMASYCSELQFLANPVTNHVQS
1 20 30 40 50 60
Intrinsically disordered region
ordered region

Figure 4. The prediction process of template-based method. The query protein is searched against a protein database with known structures to detect its homologues proteins, and then IDRs can be identified based on the Multiple Sequence Alignment (MSA). A colour version of this figure is available at BIB online: https://academic.oup.com/bib.

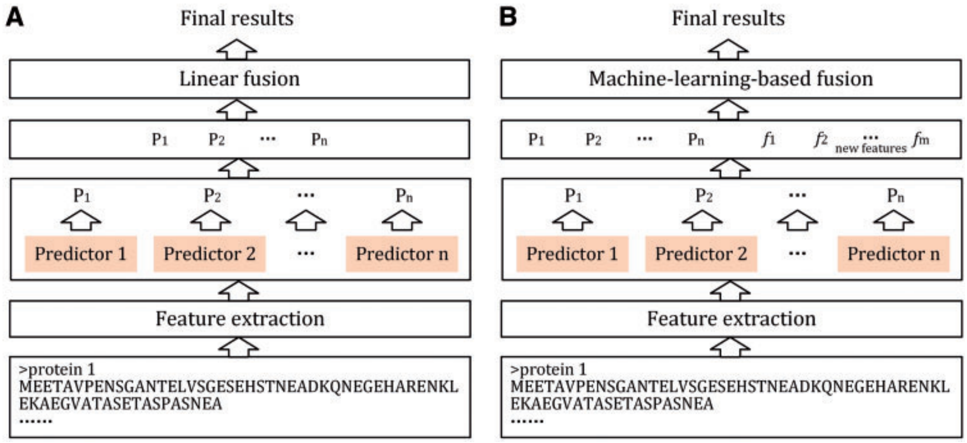
A
Final results
Linear fusion
P1 P2 ··· Pn
P1 P2 ··· Pn
Predictor 1 Predictor 2 ··· Predictor n
Feature extraction
>protein 1
MEETAVPENSGANTELVSGESEHSTNEADKQNEGEHARENKL
EKAEGVATASETASPASNEA
······
B
Final results
Machine-learning-based fusion
P1 P2 ··· Pn f1 f2 ··· new features fm
P1 P2 ··· Pn
Predictor 1 Predictor 2 ··· Predictor n
Feature extraction
>protein 1
MEETAVPENSGANTELVSGESEHSTNEADKQNEGEHARENKL
EKAEGVATASETASPASNEA
······

Figure 5. The flowchart of meta methods: (**A**) meta methods based on linear fusion, and (**B**) meta methods based on machine-learning-based fusion.

predictions. From this table, we can see that these predictors are based on different machine learning classifiers, which have impact on their performance. Generally, the sequence labeling methods, such as the CRF-based method [82] and LSTM-based method [43], outperform other approaches by incorporating the interdependence information between labels. This point will be further discussed in the performance comparison section.

Discussion

As introduced and discussed above, many computational methods of IDP/IDR prediction have been proposed. In this section, we will evaluate their predictive performance, and some open problems and perspectives will be further discussed.

Performance comparison

To evaluate the performance of different predictors, we evaluate 40 methods on the target proteins in the task of disordered region prediction in CASP10 [98]. Furthermore, we also conduct a more comprehensive comparison of 45 different predictors based on seven widely used benchmark data sets [40–45].

Some widely used performance measures are used to evaluate the performance of various methods, including AUC score, balanced accuracy (Acc) and Matthews correlation coefficient (MCC), defined as:

$$\mathrm{Acc} = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right) \tag{1}$$

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{2}$$

where *TP* (true positive) and *TN* (true negative) represent the number of correctly predicted disordered and ordered residues, respectively; *FP* (false positive) and *FN* (false negative) represent the number of misclassified ordered or disordered residues, respectively. Please note that the performance measure accuracy used in this study is a tradeoff between the specificity and sensitivity because the number of ordered residues (negative samples) is much larger than that of the disordered residues (positive samples).

Performance comparison based on CASP10

The target proteins in the task of disordered region prediction in Critical Assessment of protein Structure Prediction (CASP) have been widely used for evaluating the performance of different predictors of IDP/IDR prediction. In this article, 40 different methods are directly compared based on the disordered region prediction task in CASP10 [98]. For these unreported methods, we obtain their results by running Web servers or stand-alone programs.

The performance of various predictors based on CASP10 is shown in Table 3. According to this table, we can draw the following conclusions. (i) The deep learning models show higher performance, such as SPOT-disorder and AUCpreD. The reason is that they adopt deep hierarchical architecture, which can capture the long-range sequence information. In addition to the deep architecture, AUCpreD also adopts CRF to capture the interdependency between adjacent labels of amino acids. (ii) The performance of AUCpreD is improved when using profile information, indicating the evolutionary information in the profiles is useful for IDP/IDR prediction. (iii) Three meta-predictors (MFDp, POODLE-I and DISOPRED3) based on specialized basic predictors designed for SDR prediction or LDR prediction achieve the state-of-the-art performance, indicating that the characteristics of LDRs and SDRs are different, and they should be predicted separately. Based on these observations, we conclude that the sequence labeling predictors and meta-predictors achieve the best performance, and the classification models show better performance than physicochemical-based predictors and template-based predictors. The reason is that physicochemical-based methods only depend on the amino acid composition information, and homologous proteins cannot always be detected by the template-based methods. Compared with physicochemical-based methods and template-based methods, machine-learning-based methods can effectively use both positive and negative samples to distinguish ordered and disordered residues, and incorporate various features extracted from protein sequences, especially the sequence labeling methods are able to capture the interrelation information among all residues in proteins. The meta methods achieve the top performance because they are able to integrate various methods and features via linear fusion or machine-learning-based fusion.

Performance comparison based on seven benchmark data sets

Although the performance evaluation based on CASP10 involves 40 different predictors, some predictors without Web servers or stand-alone programs cannot be reproduced. To make a more comprehensive performance comparison among different predictors in this filed, 45 predictors are compared based on seven widely used benchmark data sets [40–45]. Following a recent review paper [39], we name the seven data sets as CASP8 [41], CASP9 [42], D494 [40], D234 [44], D25833 [45], D11925 [43] and D290 [43].

The results of the 45 methods are shown in Table 4, from which we can draw the following conclusions. (i) Similar as the results of predictors on CASP10 (see Table 3), meta-predictors and deep learning methods outperform other predictors on both small-scale data sets and large-scale data sets. (ii) Owing to the different techniques for identifying the IDPs/IDRs, and different types of disordered regions (SDRs and LDRs), predictors trained with data sets established by using one technique perform poorly on the data sets detected by other techniques. For example, the D494, D290, D234 are constructed based on PDB and DisProt with various disorder types solved by different experimental methods. The disordered proteins in CASP8 and CASP9 are solved by using X-ray or NMR techniques. IDRs in D25833 and D11925 are solved by X-ray. However, some top-performing methods achieve stable performance on different data sets, such as the meta-predictors MFDp and DISOPRED3, and deep learning predictor SPOT-disorder. The reason is that the meta-predictors predict the SDRs and the LDRs separately. Therefore, they perform well on different data sets. For the methods based on deep learning models, large data sets are often used to train the predictors, and therefore, their generalization ability is stronger than other predictors trained with small data sets.

Problems and perspectives of IDP/IDR prediction

All the aforementioned methods have made great contributions to this important field. However, there still exist some problems. In this section, we will discuss some problems in this field.

Benchmark data sets

Data sets are critical for building and evaluating the computational models. However, there are several problems in the current existing data sets of IDPs/IDRs. First, the number of experimentally verified IDPs/IDRs is still limited. IDPs/IDRs have been shown to be widely distributed in proteins [7, 8, 72], and therefore, there is a huge gap between the number of experimentally verified IDPs/IDRs and the number of existing IDPs/IDRs in nature. Therefore, more efforts should be devoted to the annotation of disorder. Benchmark data set with more samples will not only contribute to the performance improvement of existing algorithms, but also facilitate the development of new predictors. For example, deep learning techniques require a large number of training samples to avoid overfitting problem. Second, there are many biophysical techniques for detecting IDPs/IDRs [23, 24], such as NMR, X-ray, CD and SAXS, each of which provides slightly different information. Models trained with data collected by one technique cannot perform well on the data constructed by other techniques, as discussed in the 'Performance comparison' section. Therefore, collecting more IDPs/IDRs by different techniques will further improve the predictive performance. Finally, many predictors have been proposed, but most of them are trained and tested with different benchmark data sets, making it difficult for direct and objective performance comparison. To overcome these shorting comings, an updated and more comprehensive benchmark should be established containing both SDRs and LDRs identified by different techniques.

Table 2. The summary of existing Web servers and stand-alone tools for IDP/IDR prediction

Predictor	**Category**[a]	Publication year	Web sites	Classifier	Features
GlobPlot [51]	P	2003	http://globplot.embl.de/	—	The difference of amino acid propensity between 'random coil' and regular 'secondary structure'
IUPred [50]	P	2005	http://iupred.enzim.hu/	—	Amino acid composition
FoldIndex [57]	P	2005	http://bioportal.weizmann.ac.il/fldbin/findex	—	Amino acid net charge and amino acid hydrophobicity
Ucon [59]	P	2007	https://ppopen.rostlab.org/	—	Predictions of protein-specific contacts
IsUnstruct [101]	P	2011	http://bioinfo.protres.ru/IsUnstruct/	—	The energy of the ith state of a protein chain
PONDR VL-XT [53]	C	2001	http://www.pondr.com	NN	Amino acid frequency and amino acid propensities
DisEMBL [34]	C	2003	http://dis.embl.de	NN	Protein sequence
PONDR VL3 [69]	C	2005	http://www.dabi.temple.edu/disprot/predictorVSL2.php	NN	Amino acid frequency, amino acid flexibility and sequence complexity
PONDR VL3H [69]	C	2005	http://www.dabi.temple.edu/disprot/predictorVSL2.php	NN	Amino acid frequency, amino acid flexibility and sequence complexity
PONDR VL3E [69]	C	2005	http://www.dabi.temple.edu/disprot/predictorVSL2.php	NN	PSSM, amino acid flexibility and sequence complexity
RONN [52]	C	2005	https://www.strubi.ox.ac.uk/RONN	NN	Homology score between query sequence and all the prototypes
PONDR VSL2B [29]	C	2006	http://www.dabi.temple.edu/disprot/predictorVSL2.php	SVM	Amino acid frequencies, amino acid propensities and sequence complexity
PONDR VSL2P [29]	C	2006	http://www.dabi.temple.edu/disprot/predictorVSL2.php	SVM	Amino acid frequencies, amino acid propensities, sequence complexity and PSSM
Spritz [31]	C	2006	http://distill.ucd.ie/spritz/	SVM	PSSM and secondary structure predictions
PROFbval [102]	C	2006	https://ppopen.rostlab.org/	NN	Secondary structure predictions, solvent accessibility predictions, PSSM, the content in predicted regular secondary structure, the ratio of residues predicted on the surface and the protein length
Norsnet [103]	C	2007	https://ppopen.rostlab.org/	NN	Local information: PSSM, secondary structure predictions, solvent accessibility predictions, flexibility predictions; Global information: amino acid composition, the composition in predicted secondary structure and solvent accessibility, the length of the protein/domain-like fragment, the mean hydrophobicity divided by the net charge
iPDA [77]	C	2007	http://biominer.cse.yzu.edu.tw/ipda	RBFN	Condensed PSSM with respect to physicochemical properties (PSSMP)
svmPRAT [78]	C	2009	https://cs.gmu.edu/∼mlbio/svmprat/	SVM	User-supplied features
SPINE-D [30]	C	2012	http://sparks-lab.org/SPINE-D/	NN	Residue-level and window-level information calculated from amino acid composition, sequence complexity, physical parameters, PSSM, secondary structure predictions, solvent accessibility predictions and torsion-angle fluctuation predictions
SLIDER [72]	C	2014	http://biomine.cs.vcu.edu/webresults/SLIDER/	LR	Physicochemical properties, sequence complexity and amino acid composition
DisPredict [62]	C	2015	https://github.com/tamjidul/DisPredict_v1.0	SVM	Amino acids, physicochemical properties, PSSM, secondary structure predictions, accessible surface area, torsion angle fluctuation, monogram and bigram

Continued

Table 2. (continued)

Predictor	Category[a]	Publication year	Web sites	Classifier	Features
DISpro [61]	L	2005	http://www.igb.uci.edu/tools/proteomics/psss.html	RNN	PSSM, secondary structure predictions and relative solvent accessibility predictions
OnD-CRF [82]	L	2008	http://babel.ucmp.umu.se/ond-crf/	CRF	Single-sequence and secondary structure predictions
Espritz [35]	L	2012	http://protein.bio.unipd.it/espritz/	BRNN	Single-sequence or add PSSM
DeepCNF-D [86]	L	2015	http://ttic.uchicago.edu/∼wangsheng/DeepCNF_D_package_v1.00.tar.gz	CNN CRF	Amino acid-related features, evolution-related features and structure-related features
AUCpreD [87]	L	2016	http://raptorx2.uchicago.edu/StructurePropertyPred/predict/	CNN CRF	Residue-related features include amino acid identity, amino acid physic-chemical properties, amino acid propensities, correlated contact potential and reduced amino acid index. Evolution-related features include PSSM and HHM profile. Structure-level features include predicted secondary structure and solvent accessibility.
SPOT-Disorder [43]	L	2016	http://sparks-lab.org/server/SPOT-disorder/index.php	Deep bidirectional LSTM RNN	PSSM, Shannon entropy, predicted structural properties and physicochemical properties
PrDOS [54]	M	2007	http://prdos.hgc.jp/	—	Combination of a SVM-based predictor and a template-based predictor
metaPrDOS [93]	M	2008	http://prdos.hgc.jp/meta/	—	Combination of PrDOS [54], DISOPRED2 [8], DisEMBL [34], VSL2P [29], DISpro [61], IUpred [50] and POODLE-S [32]
MD [94]	M	2009	https://ppopen.rostlab.org/	—	Combination of NORSnet [103], DISOPRED2 [8], PROFbval [102], Ucon [59], IUPred [50] and some sequence properties
PONDR-FIT [3]	M	2010	http://disorder.compbio.iupui.edu/pondr-fit.php	—	Combination of PONDR VL-XT [53], PONDR VL3 [69], PONDR VSL2 [29], IUPred [50], FoldIndex [57] and TopIDP [104]
MFDp [95]	M	2010	http://biomine.cs.vcu.edu/servers/MFDp/	—	Combination of three SVM predictors specialized for SDRs, LDRs and generic IDRs. These three predictors' features are combination of IUpred [50], DISOPRED2 [8] and DISOclust [100], PSSM and other sequence properties.
CSpritz [91]	M	2011	http://protein.bio.unipd.it/cspritz/	—	Combination of Spritz [31], Punch and Espritz [35]
MetaDisorder [89]	M	2012	http://iimcb.genesilico.pl/metadisorder/	—	FloatCons: combination of 13 predictors [31, 32, 33, 50, 51, 54, 70, 74, 76, 77, 52, 105]; GSmetaDisorderMD and GSmetaDisorderMD2: combination of FloatCons and GSmetaDisorder3D [89]
MFDp2 [96]	M	2013	http://biomine.cs.vcu.edu/servers/MFDp2/	—	Combination of MFDp [95] and DisCon [99]
DisMeta [106]	M	2014	http://www-nmr.cabm.rutgers.edu/bioinformatics/disorder/	—	Combination of DisEMBL [34], DISOPRED2 [8], DISpro [61], FoldIndex [57], GlobPlot [51], IUPred [50], RONN [52] and PONDR VSL2 [29]
disCoP [44]	M	2014	http://biomine.cs.vcu.edu/servers/disCoP/	—	Combination of Espritz (the DisProt and X-ray versions) [35], Cspritz (the long version) [91], SPINE-D [30], DISOPRED2 [8], MD [94] and DISOclust [100]
DISOPRED3 [97]	M	2015	http://bioinf.cs.ucl.ac.uk/web_servers/	—	Combination of DISOPRED2 [8], a specialized predictor of LDRs and a nearest neighbor predictor
MobiDB-lite [92]	M	2017	http://protein.bio.unipd.it/mobidblite/	—	Combination of three variants of ESpritz [35], two variants of IUpred [50], DisEMBL [34] and GlobPlot [51]

[a] 'P' represents the physicochemical-based method, 'C' represents the classification model, 'L' represents the sequence labeling model and 'M' represents the meta method.

Table 3. Performance comparison of 40 different predictors on the target proteins in disordered region prediction task in CASP10

Predictor[a]		CASP 10		
		ACC	MCC	AUC
P	IsUnstruct[b] [101]	0.680	0.251	0.752
	IUPred-S[c] [50]	0.635	0.278	0.664
	IUPred-L[c] [50]	0.569	0.160	0.604
	GlobPlot[c] [51]	0.652	0.174	NA
C	SPINE-D[c] [30]	0.752	0.312	0.827
	PONDR VSL2B[b] [29]	0.701	0.241	0.744
	RONN[b] [52]	0.656	0.196	0.718
	VL3[b] [69]	0.625	0.179	0.691
	PONDR VL-XT[b] [53]	0.601	0.121	0.645
	DisEMBL-C[c] [34]	0.550	0.049	NA
	DisEMBL-R[c] [34]	0.581	0.263	NA
	DisEMBL-H[c] [34]	0.619	0.142	NA
L	AUCpreD (profile)[b] [87]	0.770	0.547	**0.913**
	SPOT-disorder [44]	0.730	**0.550**	*0.903*
	AUCpreD (no profile)[b] [87]	0.713	0.482	0.880
	ESpritz[d] [35, 98]	0.740	0.317	0.855
	OnD-CRF[e] [82]	0.727	0.311	0.814
	DISpro[b] [61, 90]	0.602	0.346	0.804
T	GSMetaDisorder3D[f] [89]	0.572	0.173	0.753
M	Prdos[g] [54, 98]	0.712	0.529	0.907
	MFDp[h] [95, 98]	0.700	0.488	0.890
	DISOPRED3 [97, 98]	0.700	0.531	0.897
	metaPrDOS[i] [93, 98]	0.778	0.385	0.879
	POODLE-I[j] [98, 107]	**0.781**	0.409	0.875
	PreDisorder[l] [90, 98]	0.769	0.400	0.873
	CASPITAv2 [98]	0.751	0.400	0.859
	disCoP[b] [44]	0.712	0.380	0.852
	GSmetaDisorderMD [89, 98]	0.727	0.341	0.844
	CSpritz[k] [91, 98]	0.759	0.316	0.829
	AIdisorder [98]	0.720	0.352	0.826
	PONDR-FIT[b] [3]	0.707	0.334	0.818
	GSmetadisorder [89, 98]	0.728	0.300	0.808
	sDisPred [98]	0.707	0.227	0.778
	GSmetaserver [89, 98]	0.699	0.204	0.778
	DisMeta [98, 106]	0.692	0.464	0.692
NA	Yang test [98]	0.738	0.376	0.872
	ZHOU-SPARKS-X [98]	0.763	0.340	0.870
	OWL2 [98]	0.686	0.387	0.821
	Slbio [98]	0.687	0.362	0.699
	Algorithmic_code [98]	0.599	0.122	0.599

The best performance is highlighted with bold font, and the performance ranked the second is highlighted with underline, and the performance ranked the third is highlighted with italic.

[a] 'P' represents the physicochemical-based method, 'C' represents the classification model, 'L' represents the sequence labeling model, 'T' represents the template-based method and 'M' represents the meta method.

[b] The results obtained from Web server.

[c] The results obtained from stand-alone package.

[d] Under group Espritz and it is a variant of ESpritz.

[e] Under group OnD-CRF2 and it is a variant of OnD-CRF.

[g] Under group Prdos-CNF and it is a variant of PrDOS.

[h] Under group biomine_dr_mixed and it is a variant of MFDp.

[i] Under group metaprdos2 and it is a variant of metaPrDOS.

[j] Under group POODLE and it is a variant of POODLE-L.

[k] Under group CSpritz and it is a variant of CSpritz.

[l] Under group MULTICOM-construct and it is variant of PreDisorder.

Feature extraction methods

Feature extraction methods are critical for IDP/IDR prediction because almost all the methods introduced in the 'Method' section are based on the features reflecting the characteristics of IDPs/IDRs. Among these available features, the profile-based features, such as PSSMs, have showed strong discriminative power (see 'Performance comparison' section). Other profile-based methods should be explored to extract the evolutionary information from the profiles, such as the Evolutionary Difference Transformation and Residue Probing Transformation [109]. Some properties used in the physicochemical-based methods can also be used as the features of the predictors based on machine learning methods, such as the residue–residue contact information used by Ucon [59]. Features reflecting the sequence order effects of amino acids are useful for studies of protein structures and functions [63], and therefore, these features and their combinations can also be applied to this field.

Machine learning algorithms

The predictors based on machine learning methods achieve state-of-the-art performance in this field. A key to improve their performance is to find a suitable machine learning algorithm. Several methods based on deep learning algorithms have been proposed to predict disordered proteins, such as SPOT-disorder [43] and AUCpreD [87], which have shown promising performance as discussed in the 'Performance comparison' section. Recently, some advanced deep learning algorithms have been proposed [110], and have been widely used in various fields, such as Deep NNs [83, 111, 112], CNNs [113–115], RNNs [116–118] and Emergent architectures [119, 120]. Based on these models, more accurate predictors can be established with suitable architectures designed for IDP/IDR prediction.

Meta-predictors

As discussed in 'Performance comparison' section, LDRs and SDRs have different characteristics. Therefore, the specialized predictors designed for SDRs and LDRs should be constructed. However, for an unseen sample, the length of its IDRs is unknown. In this regard, meta-predictors are proposed to integrate these specialized basic predictors. Recently, several methods using this strategy achieve the state-of-the-art performance, such as MFDp [95], MFDp2 [96], DisCop [44], POODLE-I [107] and DISOPRED3 [97]. The design of the basic predictors and fusion strategy are the keys to improve their performance. The basic predictors can be constructed based on different types of data sets and different machine learning models, and the fusion strategy can adopt linear fusion, machine learning techniques or other approaches. Furthermore, other properties, such as the properties of IDPs/IDRs, the properties of the motifs of IDPs/IDRs and the properties of profiles can be used to improve the predictive accuracy as introduced in a recent paper [44].

IDPs/IDRs functions

IDPs/IDRs carry out important and various functions in a cell [11, 39], for example, flexible linker, entropic bristle, assembler and activator. Recently, the first review paper [39] on IDP/IDR function prediction methods comprehensively describes various predictors in this field. These predictors [97, 121–126] have greatly facilitated the development of this field. With the

Table 4. Performance comparison of 45 different predictors based on their results on seven widely used benchmark data sets [40–45]

Predictor[a]		ACC					MCC						AUC						
		[40] D494	[41] CASP8	[42] CASP9	[44] D234	[45] D25833	[40] D494	[42] CASP9	[43] D11925	[43] D290	[44] D234	[45] D25833	[40] D494	[41] CASP8	[42] CASP9	[43] D11925	[43] D290	[44] D234	[45] D25833
P	GlobPlot [51]	0.590				0.597	0.182					0.122							0.632
	IUPred-S [50]	0.694			0.711	0.682	0.389		0.341	0.500	0.453	**0.314**	0.781			0.784	0.815	0.802	0.778
	IUPred-L [50]	0.711			0.722	0.632	0.405		0.270	0.530	0.464	0.240	0.784			0.740	0.822	0.805	0.726
	FoldIndex [57]	0.660				0.597	0.278					0.109							0.608
	Ucon [59]	0.671			0.691		0.313		0.206	0.360	0.372		0.741			0.734	0.750	0.767	
	IsUnstruct [101]			0.673				0.281							0.735				
C	DisEMBL [34]								0.327	0.390						0.789	0.760		
	DisEMBL-C [34]	0.587					0.150												
	DisEMBL-R [34]	0.626				0.674	0.323					0.308							0.787
	DisEMBL-H [34]	0.614				0.662	0.216					0.155							0.727
	DISOPRED2 [8]	0.724	0.792[b]		0.755		0.406				0.498		0.781	0.876[b]				0.828	
	PONDR VSL2B [29]	0.736			0.759	**0.742**	0.401				0.462	0.256	0.793					0.823	**0.812**
	ProfBval [102]	0.611			0.635		0.196		0.165	0.260	0.245		0.697			0.734	0.684	0.727	
	RONN [52]	0.709			0.733	0.686	0.368				0.434	0.219	0.764					0.791	0.759
	Spritz [31]	0.653					0.293												
	POODLE-S [32]				0.732						0.478							0.828	
	POODLE-L [70]				0.742						0.468							0.811	
	NORSnet [103]	0.681			0.705		0.347		0.234	0.460	0.426		0.738			0.738	0.777	0.761	
	iPDA [77]				0.710						0.465							0.841	
	svmPRAT [78]			0.736[c]				0.415[c]							0.744[c]				
	SPINE-D [30]			0.731[d]	0.777			0.365[d]	0.397	0.610	0.501				0.832[d]	0.882	0.879	0.849	
L	**DISpro** [61, 90]	0.622	0.830[e]	0.750[f]	0.648		0.318	0.365[f]		0.425	0.415		0.775	0.896[e]	0.822[f]		0.839	0.816	
	OnD-CRF [82]		0.786	0.706				0.274						0.848	0.761				
	ESpritz [35]								0.333	0.360						0.824	0.837		
	Espritz-DisProt [35]				0.779	0.541					0.525	0.110						0.855	0.731
	ESpritz-NMR [35]				0.708	0.684					0.365	0.278						0.779	0.770
	ESpritz-X-ray [35]				0.746	0.699					0.474	0.233						0.817	0.778
	SPOT-disorder [43]								0.401	**0.643**						**0.891**	**0.899**		
T	GSMetaDisorder3D [89]			0.671	0.624			0.348			0.307				0.784			0.778	
M	**PrDOS** [54]			0.754[g]	0.748			0.417[g]			0.495				**0.855[g]**			0.833	
	metaPrDOS [93]		0.760											0.871					
	MeDor [108]			0.578				0.125							0.675				
	MD [94]	0.743			0.765		0.444		0.305	0.530	0.514		**0.821**			0.813	0.841	0.847	
	PreDisorder [90]		0.809[h]											**0.918[h]**					
	POODLE-I [107]		0.794[i]	0.694[j]				0.386[j]						0.895[i]	0.831[j]				
	MFDp [95]	**0.757**		0.661[k]	**0.795**		**0.451**	0.464[k]		0.621	0.553		**0.821**		0.821[k]		0.876	0.876	

Continued

Table 4. (continued)

Predictor[a]	ACC					MCC						AUC						
	[40] D494	[41] CASP8	[42] CASP9	[44] D234	[45] D25833	[40] D494	[42] CASP9	[43] D11925	[43] D290	[44] D234	[45] D25833	[40] D494	[41] CASP8	[42] CASP9	[43] D11925	[43] D290	[44] D234	[45] D25833
PONDR-FIT [3]	0.726			0.746		0.419		0.341	0.540	0.485		0.790			0.822	0.833	0.818	
CSpritz-long [91]				0.787						0.520							0.853	
CSpritz-short [91]				0.699						0.374							0.745	
FloatCons [89]		**0.831**[l]											0.908[l]					
BinCons [89]		0.830[m]											0.897[m]					
GSmetaDisorderMD [89]			0.737	0.775			0.331			0.541				0.818			0.861	
GSmetaDisorderMD2 [89]			**0.758**[n]	0.782			0.349[n]			0.539				0.841[n]			0.858	
MetaDisorder [89]		0.802		0.790						0.549			0.830				0.861	
MFDp2 [96]									0.608							0.872		
disCoP [44]				0.786						**0.571**							**0.882**	
DISOPRED3 [97]			0.670[o]				**0.508**[o]	**0.494**	0.610					0.854[o]	0.887	0.868		

The best performance is highlighted with bold font, and the performance ranked the second is highlighted with underline and the performance ranked the third is highlighted with italic in each column. And the methods ranked in the top three in at least one measures (e.g. ACC, MCC and AUC) are highlighted with bold font.

[a] 'P' represents the physicochemical-based method, 'C' represents the classification model, 'L' represents the sequence labeling model, 'T' represents the template-based method and 'M' represents the meta method.

[b] Under group DISOPRED in CASP8.

[c] Under group Mason in CASP9.

[d] Under group Zhou-Spine-D in CASP9.

[e] Under group MULTICOM-CMFR in CASP8.

[f] Under group Multicom-refine in CASP9.

[g] Under group prdos2 in CASP 9.

[h] Under group MULTICOM in CASP8.

[i] Under group CBRC_POODLE in CASP8.

[j] Under group CBRC_Poodle in CASP9.

[k] Under group biomine_dr_pdb_c in CASP9.

[l] Under group GS-MetaServer2 in CASP8.

[m] Under group GeneSilicoMetaServer in CASP8.

[n] Under group GeneSilico in CASP9.

[o] Under group DisoPred3C in CASP9.

mature of the computational methods for IDP/IDR perdition, more efforts should be made for their function analysis.

Key Points

- Owing to the importance of studying of protein structure and function, it is critical for developing efficient and accurate computational predictors for timely identifying IDPs and IDRs.
- The databases and benchmark data sets of IDPs/IDRs are introduced and discussed, providing all the useful information of these databases.
- The computational approaches for IDP and IDR prediction are introduced, and their advantages and disadvantages are discussed. The existing Web servers and stand-alone tools in this field are given.
- The performance of existing predictors in this field is evaluated, and compared based on CASP10 and seven widely used benchmark data sets. Finally, some problems and perspectives of IDP and IDR prediction are discussed.

Acknowledgements

The authors wish to thank the four anonymous reviewers whose constructive comments were helpful to strengthen the presentation of this study.

Funding

The National Natural Science Foundation of China (No. 61672184, and 61573118), the Natural Science Foundation of Guangdong Province (2014A030313695), Guangdong Natural Science Funds for Distinguished Young Scholars (2016A030306008), Scientific Research Foundation in Shenzhen (Grant No. JCYJ20150626110425228, JCYJ20170307152201596) and Guangdong Special Support Program of Technology Young talents (2016TQ03X618).

References

1. Deng X, Eickholt J, Cheng J. A comprehensive overview of computational protein disorder prediction methods. *Mol Biosyst* 2012;**8**:114–21.
2. Deng X, Gumm J, Karki S, *et al*. An overview of practical applications of protein disorder prediction and drive for faster, more accurate predictions. *Int J Mol Sci* 2015;**16**:15384–404.
3. Xue B, Dunbrack RL, Williams RW, *et al*. PONDR-FIT: a metapredictor of intrinsically disordered amino acids. *Biochim Biophys Acta* 2010;**1804**:996–1010.
4. Wright PE, Dyson HJ. Intrinsically unstructured proteins: reassessing the protein structure-function paradigm. *J Mol Biol* 1999;**293**:321–31.
5. Galzitskaya O, Garbuzynskiy S, Lobanov MY. Prediction of natively unfolded regions in protein chains. *Mol Biol* 2006;**40**: 298–304.
6. Thomson R, Esnouf R. Prediction of natively disordered regions in proteins using a bio-basis function neural network. In: *Intelligent Data Engineering and Automated Learning - Ideal 2004, International Conference, Exeter, Uk, August 25-27, 2004, Proceedings*. 2004, p. 108–16.
7. Berman HM, Westbrook J, Feng Z, *et al*. The Protein Data Bank. *Nucleic Acids Res* 2000;**28**:235–42.
8. 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;**337**:635–45.
9. Romero P, Obradovic Z, Kissinger CR, *et al*. Thousands of proteins likely to have long disordered regions. *Pac Symp Biocomput* 1998:437–48.
10. Dunker AK, Obradovic Z, Romero P, *et al*. Intrinsic protein disorder in complete genomes. *Genome Inform Ser Workshop Genome Inform* 2000;**11**:161–71.
11. Piovesan D, Tabaro F, Micetic I, *et al*. DisProt 7.0: a major update of the database of disordered proteins. *Nucleic Acids Res* 2017;**45**:D1123–4.
12. Dunker AK, Brown CJ, Lawson JD, *et al*. Intrinsic disorder and protein function. *Biochemistry* 2002;**41**:6573–82.
13. Tompa P. The interplay between structure and function in intrinsically unstructured proteins. *FEBS Lett* 2005;**579**: 3346–54.
14. Wright PE. *Intrinsically Disordered Proteins and Their Functions*. American Association for the Advancement of Science, 2005, 432–40.
15. Iakoucheva LM, Brown CJ, Lawson JD, *et al*. Intrinsic disorder in cell-signaling and cancer-associated proteins. *J Mol Biol* 2002;**323**:573–84.
16. Midic U, Oldfield CJ, Dunker AK, *et al*. Protein disorder in the human diseasome: unfoldomics of human genetic diseases. *BMC Genomics* 2009;**10**:S12.
17. Cheng Y, Legall T, Oldfield CJ, *et al*. Abundance of intrinsic disorder in protein associated with cardiovascular disease. *Biochemistry* 2006;**45**:10448–60.
18. Raychaudhuri S, Dey S, Bhattacharyya NP, *et al*. The role of intrinsically unstructured proteins in neurodegenerative diseases. *PLoS One* 2009;**4**:e5566.
19. Uversky VN, Oldfield CJ, Dunker AK. Intrinsically disordered proteins in human diseases: introducing the D2 concept. *Annu Rev Biophys* 2008;**37**:215–46.
20. Uversky VN, Oldfield CJ, Midic U, *et al*. Unfoldomics of human diseases: linking protein intrinsic disorder with diseases. *BMC Genomics* 2009;**10 (Suppl 1)**:S7.
21. Babu MM, Van dLR, de Groot NS, *et al*. Intrinsically disordered proteins: regulation and disease. *Mol Biosyst* 2011; **21**:432–40.
22. Cheng Y, Legall T, Oldfield CJ, *et al*. Rational drug design via intrinsically disordered protein. *Trends Biotechnol* 2006;**24**:435–42.
23. Receveur-Brechot V, Bourhis JM, Uversky VN, *et al*. Assessing protein disorder and induced folding. *Proteins* 2006;**62**:24–45.
24. van der Lee R, Buljan M, Lang B, *et al*. Classification of intrinsically disordered regions and proteins. *Chem Rev* 2014;**114**: 6589–631.
25. Dosztanyi Z, Meszaros B, Simon I. Bioinformatical approaches to characterize intrinsically disordered/unstructured proteins. *Brief Bioinform* 2010;**11**:225–43.
26. He B, Wang K, Liu Y, *et al*. Predicting intrinsic disorder in proteins: an overview. *Cell Res* 2009;**19**:929–49.
27. Orosz F, Ovadi J. Proteins without 3D structure: definition, detection and beyond. *Bioinformatics* 2011;**27**:1449–54.
28. Obradovic Z, Peng K, Vucetic S, *et al*. Exploiting heterogeneous sequence properties improves prediction of protein disorder. *Proteins* 2005;**61**:176–82.
29. Peng K, Radivojac P, Vucetic S, *et al*. Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics* 2006;**7**:1.
30. Zhang T, Faraggi E, Xue B, *et al*. SPINE-D: accurate prediction of short and long disordered regions by a single neural-network based method. *J Biomol Struct Dyn* 2012;**29**:799–813.

31. 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;**34**: W164–8.
32. Shimizu K, Hirose S, Noguchi T. POODLE-S: web application for predicting protein disorder by using physicochemical features and reduced amino acid set of a position-specific scoring matrix. *Bioinformatics* 2007;**23**:2337–8.
33. Li X, Romero P, Rani M, *et al*. Predicting protein disorder for N-, C-and internal regions. *Genome Inform* 1999;**10**: 30–40.
34. Linding R, Jensen LJ, Diella F, *et al*. Protein disorder prediction: implications for structural proteomics. *Structure* 2003; **11**:1453–9.
35. Walsh I, Martin AJ, Di Domenico T, *et al*. ESpritz: accurate and fast prediction of protein disorder. *Bioinformatics* 2012; **28**:503–9.
36. Atkins JD, Boateng SY, Sorensen T, *et al*. Disorder prediction methods, their applicability to different protein targets and their usefulness for guiding experimental studies. *Int J Mol Sci* 2015;**16**:19040–54.
37. Li J, Feng Y, Wang X, *et al*. An overview of predictors for intrinsically disordered proteins over 2010-2014. *Int J Mol Sci* 2015;**16**:23446–62.
38. Meng F, Uversky V, Kurgan L. Computational prediction of intrinsic disorder in proteins. *Curr Protoc Protein Sci* 2017;**88**: 2.16.11–14.
39. Meng F, Uversky VN, Kurgan L. Comprehensive review of methods for prediction of intrinsic disorder and its molecular functions. 2017;**74**:3069–90.
40. Peng ZL, Kurgan L. Comprehensive comparative assessment of in-silico predictors of disordered regions. *Curr Protein Pept Sci* 2012;**13**:6–18.
41. Noivirt-Brik O, Prilusky J, Sussman JL. Assessment of disorder predictions in CASP8. *Proteins* 2009;**77 (Suppl 9)**:210–6.
42. Monastyrskyy B, Fidelis K, Moult J, *et al*. Evaluation of disorder predictions in CASP9. *Proteins* 2011;**79 (Suppl 10)**: 107–18.
43. Hanson J, Yang Y, Paliwal K, *et al*. Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks. *Bioinformatics* 2017;**33**:685–92.
44. Fan X, Kurgan L. Accurate prediction of disorder in protein chains with a comprehensive and empirically designed consensus. *J Biomol Struct Dyn* 2014;**32**:448–64.
45. Walsh I, Giollo M, Di Domenico T, *et al*. Comprehensive large-scale assessment of intrinsic protein disorder. *Bioinformatics* 2015;**31**:201–8.
46. Fukuchi S, Sakamoto S, Nobe Y, *et al*. IDEAL: intrinsically disordered proteins with extensive annotations and literature. *Nucleic Acids Res* 2012;**40**:D507–11.
47. Potenza E, Di Domenico T, Walsh I, *et al*. MobiDB 2.0: an improved database of intrinsically disordered and mobile proteins. *Nucleic Acids Res* 2015;**43**:D315–20.
48. Di Domenico T, Walsh I, Martin AJ, *et al*. MobiDB: a comprehensive database of intrinsic protein disorder annotations. *Bioinformatics* 2012;**28**:2080–1.
49. Oates ME, Romero P, Ishida T, *et al*. D(2)P(2): database of disordered protein predictions. *Nucleic Acids Res* 2013;**41**: D508–16.
50. 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;**21**:3433–4.
51. Linding R, Russell RB, Neduva V, *et al*. GlobPlot: Exploring protein sequences for globularity and disorder. *Nucleic Acids Research* 2003;**31**:3701–8.
52. 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;**21**:3369–76.
53. Romero P, Obradovic Z, Li X, *et al*. Sequence complexity of disordered protein. *Proteins* 2001;**42**:38–48.
54. Ishida T, Kinoshita K. PrDOS: prediction of disordered protein regions from amino acid sequence. *Nucleic Acids Res* 2007;**35**:W460–4.
55. Romero P, Obradovic Z, Kissinger C, *et al*. Identifying disordered regions in proteins from amino acid sequence. In: *International Conference on Neural Networks*, vol. **91**. 1997, 90–5.
56. Uversky VN, Gillespie JR, Fink AL. Why are “natively unfolded” proteins unstructured under physiologic conditions? *Proteins* 2000;**41**:415–27.
57. 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;**21**:3435–8.
58. Galzitskaya OV, Garbuzynskiy SO, Lobanov MY. FoldUnfold: web server for the prediction of disordered regions in protein chain. *Bioinformatics* 2006;**22**:2948–9.
59. Schlessinger A, Punta M, Rost B. Natively unstructured regions in proteins identified from contact predictions. *Bioinformatics* 2007;**23**:2376–84.
60. 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;**347**:827–39.
61. Cheng J, Sweredoski MJ, Baldi P. Accurate prediction of protein disordered regions by mining protein structure data. *Data Min Knowl Discov* 2005;**11**:213–22.
62. Iqbal S, Hoque MT. DisPredict: a predictor of disordered protein using optimized RBF kernel. *PLoS One* 2015;**10**:e0141551.
63. Liu B, Liu F, Wang X, *et al*. Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Res* 2015;**43**:W65–71.
64. Wei L, Zou Q. Recent progresses in machine learning-based methods for protein fold recognition. *Int J Mol Sci* 2016;**17**: 2118.
65. Li D, Ju Y, Zou Q. Protein folds prediction with hierarchical structured SVM. *Curr Proteom* 2016;**13**:79–85.
66. Lin H, Liang ZY, Tang H, *et al*. Identifying sigma70 promoters with novel pseudo nucleotide composition. *IEEE/ACM Trans Comput Biol Bioinform* 2017, in press.
67. Zhang CJ, Tang H, Li WC, *et al*. iOri-Human: identify human origin of replication by incorporating dinucleotide physicochemical properties into pseudo nucleotide composition. *Oncotarget* 2016;**7**:69783–93.
68. Yang H, Tang H, Chen XX, *et al*. Identification of secretory proteins in mycobacterium tuberculosis using pseudo amino acid composition. *biomed Res Int* 2016;**2016**:5413903.
69. Peng K, Vucetic S, Radivojac P, *et al*. Optimizing long intrinsic disorder predictors with protein evolutionary information. *J Bioinform Comput Biol* 2005;3:35–60.
70. Hirose S, Shimizu K, Kanai S, *et al*. POODLE-L: a two-level SVM prediction system for reliably predicting long disordered regions. *Bioinformatics* 2007;**23**:2046–53.
71. Han P, Zhang X, Feng ZP. Predicting disordered regions in proteins using the profiles of amino acid indices. *BMC Bioinformatics* 2009;**10 (Suppl 1)**:S42.

72. Peng Z, Mizianty MJ, Kurgan L. Genome-scale prediction of proteins with long intrinsically disordered regions. *Proteins* 2014;**82**:145–58.
73. Pollastri G, McLysaght A. Porter: a new, accurate server for protein secondary structure prediction. *Bioinformatics* 2005; **21**:1719–20.
74. Ward JJ, McGuffin LJ, Bryson K, *et al*. The DISOPRED server for the prediction of protein disorder. *Bioinformatics* 2004;**20**: 2138–9.
75. Jones DT, Ward JJ. Prediction of disordered regions in proteins from position specific score matrices. *Proteins* 2003;**53**:573–8.
76. Su CT, Chen CY, Ou YY. Protein disorder prediction by condensed PSSM considering propensity for order or disorder. *BMC Bioinformatics* 2006;**7**:319.
77. Su CT, Chen CY, Hsu CM. iPDA: integrated protein disorder analyzer. *Nucleic Acids Res* 2007;**35**:W465–72.
78. Rangwala H, Kauffman C, Karypis G. svmPRAT: SVM-based protein residue annotation toolkit. *BMC Bioinformatics* 2009; **10**:439.
79. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* 1999;**292**: 195–202.
80. Altschul SF, Madden TL, Schaffer AA, *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;**25**:3389–402.
81. Hecker J, Yang JY, Cheng J. Protein disorder prediction at multiple levels of sensitivity and specificity. *BMC Genomics* 2008;**9 (Suppl 1)**:S9.
82. Wang L, Sauer UH. OnD-CRF: predicting order and disorder in proteins using [corrected] conditional random fields. *Bioinformatics* 2008;**24**:1401–2.
83. Heffernan R, Paliwal K, Lyons J, *et al*. Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. *Sci Rep* 2015;**5**:11476.
84. Heffernan R, Dehzangi A, Lyons J, *et al*. Highly accurate sequence-based prediction of half-sphere exposures of amino acid residues in proteins. *Bioinformatics* 2016;**32**: 843–9.
85. Paliwal K, Lyons J, Heffernan R. A short review of deep learning neural networks in protein structure prediction problems. *Adv Techn Biol Med* 2015;**189**:54–63.
86. Wang S, Weng S, Ma J, *et al*. DeepCNF-D: predicting protein order/disorder regions by weighted deep convolutional neural fields. *Int J Mol Sci* 2015;**16**:17315–30.
87. Wang S, Ma J, Xu J. AUCpreD: proteome-level protein disorder prediction by AUC-maximized deep convolutional neural fields. *Bioinformatics* 2016;**32**:i672–9.
88. Liu B, Chen J, Wang X. Application of learning to rank to protein remote homology detection. *Bioinformatics* 2015;**31**:3492–8.
89. Kozlowski LP, Bujnicki JM. MetaDisorder: a meta-server for the prediction of intrinsic disorder in proteins. *BMC Bioinformatics* 2012;**13**:111.
90. Deng X, Eickholt J, Cheng J. PreDisorder: ab initio sequence-based prediction of protein disordered regions. *BMC Bioinformatics* 2009;**10**:436.
91. Walsh I, Martin AJ, Di Domenico T, *et al*. CSpritz: accurate prediction of protein disorder segments with annotation for homology, secondary structure and linear motifs. *Nucleic Acids Res* 2011;**39**:W190–6.
92. Necci M, Piovesan D, Dosztanyi Z, *et al*. MobiDB-lite: Fast and highly specific consensus prediction of intrinsic disorder in proteins. *Bioinformatics* 2017, in press.
93. Ishida T, Kinoshita K. Prediction of disordered regions in proteins based on the meta approach. *Bioinformatics* 2008;**24**: 1344–8.
94. Schlessinger A, Punta M, Yachdav G, *et al*. Improved disorder prediction by combination of orthogonal approaches. *PLoS One* 2009;**4**:e4433.
95. Mizianty MJ, Stach W, Chen K, *et al*. Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources. *Bioinformatics* 2010;**26**: i489–96.
96. Mizianty MJ, Peng Z, Kurgan L. MFDp2: Accurate predictor of disorder in proteins by fusion of disorder probabilities, content and profiles. *Intrinsically Disord Proteins* 2013;**1**:e24428.
97. Jones DT, Cozzetto D. DISOPRED3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics* 2015;**31**:857–63.
98. Monastyrskyy B, Kryshtafovych A, Moult J, *et al*. Assessment of protein disorder region predictions in CASP10. *Proteins* 2014;**82 (Suppl 2)**:127–37.
99. Mizianty MJ, Zhang T, Xue B, *et al*. In-silico prediction of disorder content using hybrid sequence representation. *BMC Bioinformatics* 2011;**12**:245.
100. McGuffin LJ. Intrinsic disorder prediction from the analysis of multiple protein fold recognition models. *Bioinformatics* 2008;**24**:1798–804.
101. Lobanov MY, Galzitskaya OV. The Ising model for prediction of disordered residues from protein sequence alone. *Phys Biol* 2011;**8**:035004.
102. Schlessinger A, Yachdav G, Rost B. PROFbval: predict flexible and rigid residues in proteins. *Bioinformatics* 2006;**22**:891–3.
103. Schlessinger A, Liu J, Rost B. Natively unstructured loops differ from other loops. *PLoS Comput Biol* 2007;**3**:e140.
104. Campen A, Williams RM, Brown CJ, *et al*. TOP-IDP-scale: a new amino acid scale measuring propensity for intrinsic disorder. *Protein Pept Lett* 2008;**15**:956–63.
105. Medina MW, Gao F, Naidoo D, *et al*. Coordinately regulated alternative splicing of genes involved in cholesterol biosynthesis and uptake. *PLoS One* 2011;**6**:e19420.
106. Huang YJ, Acton TB, Montelione GT. DisMeta: a meta server for construct design and optimization. *Methods Mol Biol* 2014; **1091**:3–16.
107. Hirose S, Shimizu K, Noguchi T. POODLE-I: disordered region prediction by integrating POODLE series and structural information predictors based on a workflow approach. *In Silico Biol* 2010;**10**:185–91.
108. Lieutaud P, Canard B, Longhi S. MeDor: a metaserver for predicting protein disorder. *BMC Genomics* 2008;**9 (Suppl 2)**: S25.
109. Zhang J, Liu B. PSFM-DBT: identifying DNA-binding proteins by combing position specific frequency matrix and distance-bigram transformation. *Int J Mol Sci* 2017;**18**:1856.
110. Min S, Lee B, Yoon S. Deep learning in bioinformatics. *Brief Bioinform* 2016, in press.
111. Spencer M, Eickholt J, Jianlin C. A deep learning network approach to AB initio protein secondary structure prediction. *IEEE/ACM Trans Comput Biol Bioinform* 2015;**12**:103–12.
112. Lyons J, Dehzangi A, Heffernan R, *et al*. Predicting backbone Calpha angles and dihedrals from protein sequences by stacked sparse auto-encoder deep neural network. *J Comput Chem* 2014;**35**:2040–6.
113. Alipanahi B, Delong A, Weirauch MT, *et al*. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol* 2015;**33**:831–8.

114. Lanchantin J, Singh R, Wang B, *et al*. Deep motif dashboard: visualizing and understanding genomic sequences using deep neural networks. *Pac Symp Biocomput* 2016;**22**: 254–65.
115. Zeng H, Edwards MD, Liu G, *et al*. Convolutional neural network architectures for predicting DNA-protein binding. *Bioinformatics* 2016;**32**:i121–7.
116. Baldi P, Brunak S, Frasconi P, *et al*. Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics* 1999;**15**:937–46.
117. Baldi P, Pollastri G, Andersen CA, *et al*. Matching protein beta-sheet partners by feedforward and recurrent neural networks. *Proc Int Conf Intell Syst Mol Biol* 2000;**8**:25–36.
118. Heffernan R, Yang Y, Paliwal K, *et al*. Capturing non-local interactions by long short term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers, and solvent accessibility. *Bioinformatics* 2017, in press.
119. Di Lena P, Nagata K, Baldi P. Deep architectures for protein contact map prediction. *Bioinformatics* 2012;**28**:2449–57.
120. Baldi P, Pollastri G. The principled design of large-scale recursive neural network architectures-DAG-RNNs and the protein structure prediction problem. *J Mach Learn Res* 2003; **4**:575–602.
121. Meng F, Na I, Kurgan L, *et al*. Compartmentalization and functionality of nuclear disorder: intrinsic disorder and protein-protein interactions in intra-nuclear compartments. *Int J Mol Sci* 2015;**17**.
122. Malhis N, Gsponer J. Computational identification of MoRFs in protein sequences. *Bioinformatics* 2015;**31**:1738–44.
123. Meng F, Kurgan L. DFLpred: High-throughput prediction of disordered flexible linker regions in protein sequences. *Bioinformatics* 2016;**32**:i341–50.
124. Peng Z, Kurgan L. High-throughput prediction of RNA, DNA and protein binding regions mediated by intrinsic disorder. *Nucleic Acids Res* 2015;**43**:e121.
125. Yan J, Dunker AK, Uversky VN, *et al*. Molecular recognition features (MoRFs) in three domains of life. *Mol Biosyst* 2016; **12**:697–710.
126. Malhis N, Jacobson M, Gsponer J. MoRFchibi SYSTEM: software tools for the identification of MoRFs in protein sequences. *Nucleic Acids Res* 2016;**44**:W488–93.
127. Ghalwash MF, Dunker AK, Obradovic Z. Uncertainty analysis in protein disorder prediction. *Mol Biosyst* 2012;**8**:331–391.