Cell Research (2009) 19:929-949.

www.nature.com/cr

REVIEW

Predicting intrinsic disorder in proteins: an overview

Bo He[1], Kejun Wang[1], Yunlong Liu[2, 5, 6], Bin Xue[2, 3, 4], Vladimir N Uversky[2, 3, 4, 7], A Keith Dunker[2, 3, 4]

[1]College of Automation, Harbin Engineering University, Harbin, Heilongjiang 150001, China; [2]Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, IN 46202, USA; [3]Department of Biochemistry and Molecular Biology, Indiana University School of Medicine, Indianapolis, IN 46202, USA; [4]Institute for Intrinsically Disordered Protein Research, Indiana University School of Medicine, Indianapolis, IN 46202, USA; [5]Division of Biostatistics, Department of Medicine, Indiana University School of Medicine, Indianapolis, IN 46202, USA; [6]Center for Medical Genomics, Indiana University School of Medicine, Indianapolis, IN 46202, USA; [7]Institute for Biological Instrumentation, Russian Academy of Sciences, 142290 Pushchino, Moscow Region, Russia

The discovery of intrinsically disordered proteins (IDP) (i.e., biologically active proteins that do not possess stable secondary and/or tertiary structures) came as an unexpected surprise, as the existence of such proteins is in contradiction to the traditional "sequence→structure→function" paradigm. Accurate prediction of a protein's predisposition to be intrinsically disordered is a necessary prerequisite for the further understanding of principles and mechanisms of protein folding and function, and is a key for the elaboration of a new structural and functional hierarchy of proteins. Therefore, prediction of IDPs has attracted the attention of many researchers, and a number of prediction tools have been developed. Predictions of disorder, in turn, are playing major roles in directing laboratory experiments that are leading to the discovery of ever more disordered proteins, and thereby leading to a positive feedback loop in the investigation of these proteins. In this review of algorithms for intrinsic disorder prediction, the basic concepts of various prediction methods for IDPs are summarized, the strengths and shortcomings of many of the methods are analyzed, and the difficulties and directions of future development of IDP prediction techniques are discussed.

***Keywords*:** protein, intrinsic disorder, prediction method

Cell Research (2009) **19**:929-949. doi: 10.1038/cr.2009.87; published online 14 July 2009

Introduction

The traditional "sequence → structure → function" model for describing protein activity states that the amino acid sequence determines the higher structures of a protein molecule, including its secondary and tertiary conformations, as well as quaternary complexes and further states that the formation of a definite ordered structure represents the foundation for the function of the protein. If particular conditions, such as acid, urea, or high temperature, cause a protein to lose its unique and ordered structure, then it loses its ability to carry out function and is considered to have become denatured. The first statement of this concept that denaturation arises from loss of structure was made by Hsien Wu almost eight decades ago [1, 2].

For more than five decades, researchers have been discovering individual proteins that possess no definite ordered three-dimensional structure but still play important biological roles. The discovery rate for such proteins has been increasing continually and has become especially rapid during the last decade [3]. The discovery and characterization of these proteins is becoming one of the fastest growing areas of protein science.

Many such proteins with no unique structure are involved in key biological processes including cell cycle control, regulation, recognition, and signaling [4-8]. Some researchers believe that structural flexibility and plasticity originating from the lack of a definite ordered three-dimensional structure represents a major functional advantage for these proteins. These proteins are able to interact with and bind to a broad range of ligands, includ-

Correspondence: Kejun Wang[a], A Keith Dunker[b]
[a]Tel: +86-451-8256-9906; Fax: +86-451-8256-9906
[a]E-mail: wangkejun@hrbeu.edu.cn
[b]E-mail: kedunker@iupui.edu

ing partners such as themselves or other proteins, membranes, and nucleic acids [9-12]. These binding events typically involve coupled binding and folding [13]. An alternative idea was that binding of disordered regions depends on conformational selection from the structural ensemble [14] perhaps via "pre-formed elements" [15] that dominate the ensemble. Recently, it has been suggested that protein-protein interactions actually involve a combination of coupled binding and folding and conformational selection [16]. With regard to protein-protein interactions, papers have appeared recently that follow the structural transitions as these highly flexible regions associate with their partners [17]. Further study shows that the same flexible region of a given protein can fold differently when binding different partners and also that different amino acid sequences can use their flexibility to fold onto a common binding site on the same protein partner [11] just as predicted more than a decade ago [18].

In addition to being rich in binding sites for various partners, these regions that lack stable, specific three-dimensional structure have also been found to be important loci for alternative splicing [19] and for enzyme-driven posttranslational modifications such as phosphorylation, methylation, or acetylation [8]. Since regions lacking structure are rich in binding sites, and since these binding sites can be readily modulated or eliminated both by posttranslational modification and by alternative splicing, such regions that don't form structure are becoming increasingly viewed as crucial both for signaling in unicellular eukaryotic organisms and for signaling diversity in multicellular organisms [3].

For many examples, the given disordered regions are not known to bind to any partner, but they still carry out important functions such as providing flexible linkers between structured domains or providing flexible tails that regulate the structured domains [4, 20, 21]. One recent study showed that a particular flexible linker maintains its length and flexibility across divergent species despite having little or no obvious amino acid sequence conservation. Perhaps for this linker, the amino acid composition is more important than the specific details of the sequence [22].

Various researchers have used different terms to describe these proteins and regions, including intrinsically disordered [23], intrinsically unstructured [20, 24], natively unfolded [21, 25], natively disordered proteins [26], and highly flexible [27, 28]. Some of these proteins and regions have been shown to contain partial or transient secondary and/or tertiary structural organization [29, 30]. Our view is that the terms "unfolded" and "unstructured" would be misleading for these partially structured examples, and "flexible" has been used to describe structured regions with high B-factors rather than regions lacking structure, so herein and elsewhere we use the term intrinsically disordered protein (IDP) to cover the wide range of possibilities.

Figure 1 provides a schematic view of calcineurin, a protein that uses a highly flexible region for signaling and regulation [23, 31]. This protein is a serine/threonine protein phosphatase that contains a catalytic A subunit (which is a calcium/calmodulin-activated serine/threonine phosphatase), and a calmodulin-like B subunit. In addition, this protein complex contains four disordered regions, labeled first to fourth, in the order of decreasing length. The lengths of these regions are 95, 35, 13, and 4 residues, respectively, as determined by missing electron density in the crystal structures [31]. A 19-residue autoinhibitory peptide lies between the first and second disordered regions. At low calcium, this autoinhibitory peptide is bound to the active site, and the enzyme activity is turned off. In a signaling event, increased calcium leads to activation of calmodulin, which binds to many target proteins. Calcineurin is among those proteins that have a calmodulin-binding site, and this site is located in the first disordered region, as shown in Figure 1. This region probably becomes helical upon calmodulin binding, and the binding leads to displacement of the autoinhibitory peptide, and thereby turns on calcineurin's serine/threonine phosphatase activity.

Calmodulin surrounds its target upon binding, so locating its binding site in calcineurin within a region of disorder makes the target accessible all the way around as is required. Therefore, the region of disorder appears to be essential to the regulation of calcineurin by calcium/calmodulin [23, 31]. Indeed, when experiments have been performed on many different calmodulin-activated enzymes, the results usually indicate that calmodulin's binding target is located within a trypsin-sensitive region [32] that is very likely to be intrinsically disordered. Note that calcineurin connects two major signaling pathways, calcium/calmodulin signaling with phosphorylation/dephosphorylation signaling. Not surprisingly, therefore, this protein and its relatives play important roles in many tissue types and in a wide variety of eukaryotic organisms.

To account for the constantly increasing number of experimentally characterized IDPs, a databank for IDPs, named DisProt [33], has been built and is being continually updated. DisProt introduces 32 kinds of IDP-related functional subclasses. Other work indicates that 238 functions given in SwissProt are associated with IDPs [8], but so far DisProt has not been updated to include these new results. Figure 2 gives the currently observed number of IDPs for each functional subclass in DisProt. The

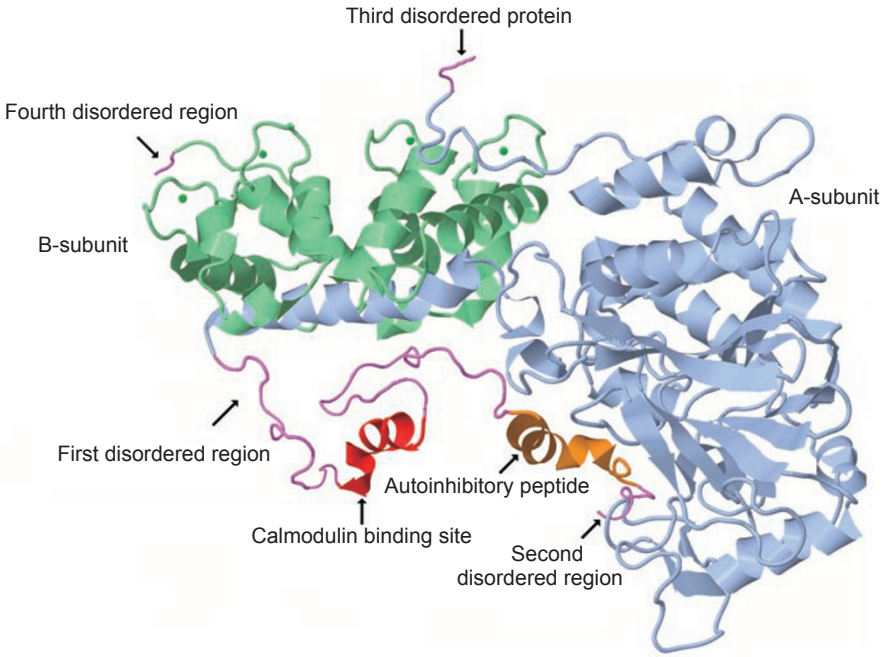

Figure 1 Structure of calcineurin with essential disorder. The A subunit of calcineurin contains a phosphatase domain (blue), a helix (blue) that binds the B subunit (green), a calmodulin-binding target (red), and an autoinhibitory peptide (saffron) that binds to the active site on the phosphatase domain. The B subunit (green) resembles calmodulin. The complex also contains four disordered regions (pink). The first disordered region (95 amino acids), connects the end of a helix (residue 373) to the autoinhibitory peptide (residue 469) and contains a helical calmodulin-binding site; the second disordered region (35 amino acids) follows the other end of the autoinhibitory peptide (residue 486); the third disordered region (13 amino acids) is located at the amino-terminus of the A subunit, and the fourth disordered region (4 amino acids) is joined to the B subunit (residue 5). The calmodulin-binding target and the autoinhibitory peptide likely lose their structures when not bound to their partners, and thus probably utilize coupled binding and folding mechanisms as have been shown for disordered regions in other proteins. These binding segments are shown to be structured here in the absence of their partners in order to indicate their approximate locations in the disordered regions.

biological functions of many IDPs are still unknown. Because the DisProt entries are randomly acquired, Figure 2 shouldn't be used as an indicator of the natural frequency of various functions attributed to disordered proteins. In fact, this figure only reflects the abundance of functional categories based upon the proteins currently annotated in DisProt.

Based on a very small number of proteins, Williams [34] suggested an approach for using amino acid sequence for identifying proteins that form random coils rather than globular structures, but this approach was never carefully tested. Later, Dunker and Uversky and their coworkers independently published the first well-tested predictors of IDPs [35, 36]. Since then, numerous researchers have designed many algorithms to predict disordered proteins utilizing specific biochemical properties and biased amino acid compositions of IDPs. Various prediction ideas and different computing techniques have been utilized. Many of these predictors including PONDR®s [36-41], FoldIndex [42], GlobPlot [43], DisEMBL [44], DISOPRED and DISOPRED2 [45-48], DRIPPRED [49], IUPred [50, 51], FoldUnfold [52-54], RONN [55], DISpro [56], DisPSSMP and DisPSSMP2 [57, 58], Spritz [59] and PrDOS [60], etc. can be accessed via public servers and evaluate intrinsic disorder on a per-residue basis. Since the first predictors were published, more than 50 predictors of disorder have been developed, with their accumulation over time shown in Figure 3. The legend to this figure contains references to all of the disorder predictors we have been able to find so far. In the text below, we discuss many but not all of these predictors, where the omissions and limitations in our discussions have mostly to do with time and space.

Most of the published predictors are similar in the prediction of long disordered regions, but they do differ significantly in the local details of the outputs. There are also binary disorder predictors, *e.g.*, the charge-hydropathy (CH)-plot [35] and the cumulative distribution

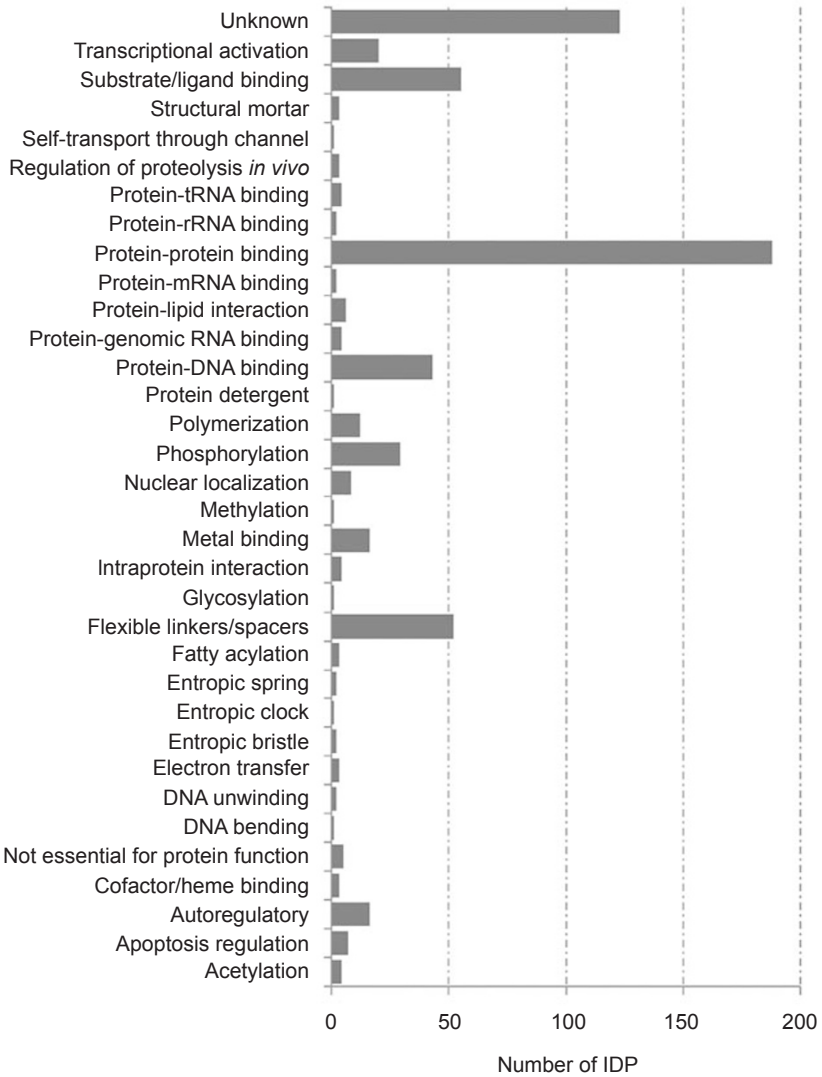

Figure 2 The number of IDPs versus each functional subclass. The DisProt database introduces 32 kinds of IDP-related functional subclasses. The x axis gives the number of IDPs involved in each subclass. Currently, there are 123 IDPs in the DisProt database whose functions are still unknown. Intrinsic disorder in a given protein can be found by multiple experimental and computational techniques. The experimental techniques include X-ray diffraction, circular dichroism, nuclear magnetic resonance spectroscopy, intrinsic and extrinsic fluorescence, dynamic light scattering, small angle X-ray scattering, gel-filtration, infrared spectroscopy, Raman optical activity and Raman spectroscopy, limited proteolysis, and many others [121]. Many of these techniques are costly and require both a lot of time and an extensive expertise. Furthermore, they are not easily applicable for large-scale studies, *e.g.*, for a proteome-wide analysis.

function (CDF) analysis [61], both of which evaluate the probability that an entire protein is structured or disordered. Links to many of these predictors can be found in the Disordered Protein Database (http://www.DisProt.org) [33]. Links are under construction for many of the predictors that are currently missing from DisProt.

From the biologist's point of view, what is especially interesting is that in many cases predictions of protein disorder are being used to guide laboratory experiments [62-64], which are in turn leading to the discovery of increasing numbers of disordered proteins. We are witnessing the development of a positive feedback loop involving prediction-experiment-prediction, etc., and this loop is leading to further increases in the rates of discovery for IDPs. To illustrate the growing interaction between disorder prediction and laboratory experiments, seven interesting examples are presented below.

In the first example, disorder predictions were instru-

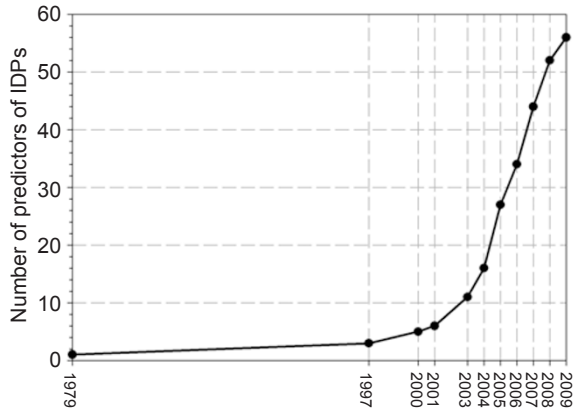

Figure 3 The total number of IDP predictors. The list of predictors includes the following: the first suggested predictor of IDPs [34]; the first formal predictor of IDPs [36]; predictor of ID in calcineurin family [89]; CH-Plot [35]; CDF [90]; PONDR® VL-XT [38]; GlobPlot [43]; DisEMBL [44]; DISOPRED [45]; flavors of protein disorder [119]; NORSp [122]; predictor by using reduced amino acid alphabet [91]; DISOPRED2 [46]; DRIPPRED [49]; FoldUnfold [52, 53]; Softberry (http://www.softberry.com); VaZyMolO [97]; PONDR® VL3-E [39]; IUPred [50, 51]; FoldIndex [42]; RONN [55]; DISpro [56]; PONDR® VSL1 [40]; CDF [61]; combined CDF/CH-Plot predictor [61]; α-MoRF [88]; Prelink [115]; PONDR® VSL2 [41]; Spritz [59]; DisPSSMP [57]; IUP predictor [103]; disorder prediction in calmodulin partners [32]; Decision trees [68]; Wiggle [123]; iPDA [58]; PrDOS [60]; SGT [105]; Ucon [104]; α-MoRF II [87]; composition profiler [124]; POODLE-L [102]; POODLE-S [101]; POODLE-W [105]; NORSnet [125]; OnD-CRF [107]; predictor by using bayesian multinomial classifier [106]; DISOclust [111]; Top-IDP [126]; DPROT [127]; hierarchical classifier [128]; MetaPrDOS [109]; MeDor [112]; Draai [129]; CDF-ALL [108];and IUPforest-L [130].

mental in guiding the structural and functional analyses of the measles virus nucleoprotein (N) [62, 65, 66]. In this example, the C-terminal domain of N (aa 401-525, N_{TAIL}) was predicted and then experimentally shown to be intrinsically disordered [65]. Next, a predicted α-helical molecular recognition feature, α-MoRF, was predicted in the N_{TAIL} region, which was then shown to bind to measles virus phosphoprotein (P) [66]. A truncated N_{TAIL} lacking the region containing the predicted α-MoRF failed to bind P, thus showing that the predicted α-MoRF region is likely required for binding to P [66]. The predicted binding site (residues 486-499) was confirmed to be an almost perfect match to the actual binding site (residues 484-504) as determined by X-ray crystallography [67].

In the second example, disorder predictions helped to guide experiments showing that the organizing domain of RNase H is disordered. Furthermore, the predictions also helped to identify several binding sites for several of the partners that join the complex [64]. Finally, a binding site initially identified by disorder prediction was crystallized with its partner and the structure of the complex was determined. In this case, the predictor suggested residues 834-851 as the binding segment, whereas the three-dimensional structure contained a structured segment that included the 833-847 segment, [68] indicating quite a good match between the predicted and observed binding segment. Further investigation of the structure of the complex suggests the need for the deletion of two to three residues from the amino-terminus and the addition of approximately three residues on the carboxyl side to more closely match the extent of the binding groove, but even without this deletion and addition, the agreement cited above is already excellent.

In the third example, disorder predictions helped to define studies showing that alternative splicing of a disordered region of a hox protein plays a key role in the development of bilateral symmetry in organisms ranging from drosophila to humans [69]. Work in progress shows that the DNA-binding affinities to a number of different DNA targets become altered as a result of the alternative splicing in the disordered region. The authors suggest that the variously spliced disordered region may act as an antenna that integrates tissue-specific information to direct the hox protein to the correct DNA-binding sites (Bondos, personal communication).

In the fourth example, disorder prediction and analysis showed that a flexible linker in replication protein A showed an extremely high sequence variability over evolutionary time as compared to the structured domains of the same protein [70]. This was followed up by NMR studies of this linker region from a divergent set of species, with the finding that the flexibility remained almost constant despite the nearly complete absence of sequence conservation [22].

In the fifth example, disorder predictions guided investigations of protein regulation in *Saccharomyces cerevisiae*. The most highly disordered proteins were found to be the most tightly regulated, and these highly regulated disordered proteins are generally associated with signaling and posttranslational modification [71, 72].

In the sixth example, disorder prediction guided the successful crystallization of NEIL1, a human homolog of *Escherichia coli* DNA glycosylase endonuclease VIII by indicating a disordered region that likely prevented crystallization. Crystallization and structure determination were accomplished by using genetic engineering to remove the predicted region of disorder [73].

In the seventh and last example discussed here, disorder prediction and experience with disordered proteins

are guiding the development of novel proteomics methods that are identifying proteins previously missed by standard approaches [74-77]. An interesting aspect of this recent work has been the development of additional evidence for a connection between alternative splicing and intrinsic disorder [78].

The seven examples given above illustrate a few of the many ways by which experimentalists are making use of disorder predictions to help design experimental approaches. This list of seven was selected to be illustrative and is not by any means comprehensive.

In this review, basic concepts and ideas for disorder prediction by various algorithms are analyzed, the difficulties of the IDP prediction are illuminated, and some likely future trends in the development of the IDP prediction techniques are discussed. Our goal here is to help increase the intelligent use of these predictors by molecular biologists working at the bench.

The current state of IDP prediction

Recent developments regarding disorder prediction

Recently, the studies on IDPs have gained significant attention resulting in a rapid growth of the number of research articles and reviews. Since the recent reviews mainly discuss IDP predictors that were designed before 2005 [79-81], here we are focusing on more recent developments and are categorizing the IDP prediction methods in terms of their key concepts and ideas.

An important development has been the inclusion of disorder identification in the Critical Assessment of Structure Prediction (CASP) meetings [82]. Participants in these meetings make predictions on amino acid sequences as the structures of these proteins are being determined, but before the structures are known. Once the structure of a given protein is completed, structure predictions on that protein are cut off. An independent group of researchers then compares the various predictions from many research groups with the observed structures. In this way, the predictions are blind, and the third party evaluations are unbiased.

The evaluations of the disorder predictions for CASP5, 6, and 7 have been published [83-85]. These evaluations provide useful insight into the various predictors of disorder. With regard to CASP7, nineteen different predictors were evaluated and the best five achieved overall accuracies of ~ 69 % to ~ 78% in terms of the averaged value of specificity and sensitivity for a two state prediction. In other words, the accuracy was estimated as percent correct on disordered regions plus percent correct on structured regions divided by two, regardless of the degree of imbalance between the number of residues in disordered and structured regions. These predictors also gave areas under their receiver operating characteristics curves (ROC curves) ranging from 0.822 ± 0.008 to 0.860 ± 0.007, where the ROC curve is a plot of the true-positive rate versus the false-positive rate for a given predictor. A random predictor would give a value of 0.5 for the area under the ROC curve, and a perfect predictor would give 1.0. Thus, the observed values, which are > 0.80, indicate fairly good predictors. Although the predictors could be ranked by their accuracies or by their areas under their ROC curves, the total number of predictions at each CASP exercise is rather small, so the CASP results should not be used to claim that one predictor is more accurate than another.

In our view, the current limitation to further improvement in prediction accuracies comes from noise in the structured and disordered protein data. Unstructured proteins can form complexes that become structured [86], and unless care is taken when selecting structured proteins from the Protein Data Bank, structured proteins arising from disorder would contribute noise to the training set for structured protein. Likewise, regions that are characterized as disordered experimentally can undergo coupled binding and folding. Indeed, PONDR® VL-XT has proved to be useful in predicting disordered regions that bind to protein partners [64, 87, 88]. Such regions often have strong structure-forming tendencies for a localized region of sequence, and so such regions could provide noise for the training set for disordered protein.

An IDP prediction is built on the basis of the analysis of some collection of protein properties. By now, there have been many predictors. For many of these predictors, the name, a brief cognate description, and the corresponding references are listed in Table 1. Below is an historical overview of the development of disorder predictors. We first provide an overall representation of PONDR®s, which are a series of predictors that have various versions, each with its own specificity. Next, we discuss the development of some other predictors. We also elaborate on the likely characteristics and biological significance of IDPs demonstrated by these predictors.

The development of disorder predictors from a historical view

The significance of IDPs has led to an increase in the number of IDP predictors. Besides PONDR®, there are many additional IDP predictors, some of which are undergoing continual modification and improvement. In order to provide an historical perspective, we arrange these predictors based on the time of publication of the original predictor and show the accumulated number of published predictors over time in Figure 3. This arrangement suggests that the development of these predictors

Table 1 Predictors of IDPs

Predictor	Publication year	Brief description
PONDR® [36-41] http://www.pondr.com	1997-2006	PONDR®s include several predictors that predict isordered regions with different length or in any location of a sequence. All PONDR® predictors exhibit reasonably good performance.
GlobPlot [43] http://globplot.embl.de	2003	The key idea of GlobPlot is the relative propensity of an amino acid residue to be in an ordered or disordered state.
DisEMBL [44] http://dis.embl.de	2003	DisEMBL is able to predict three kinds of disordered structure, including loops/coils, hot loops, and those that are missing from the PDB X-ray structures.
DISOPRED [45]	2003	DISOPRED applies neural networks to inputs of whole sequence information.
DISOPRED2 [46] http://bioinf.cs.ucl.ac.uk/disopred	2004	DISOPRED2 directly trains on the whole sequence by using SVM.
Weather's method [91]	2004	Weather's method uses SVM analysis of a linear combination of composition vectors.
DRIPPRED [49] http://www.sbc.su.se/~maccallr/disorder/	2004	DRIPPRED is based on Kohonen's self-organizing map and received a good evaluation at CASP6.
FoldUnfold [52-54] http://skuld.protres.ru/~mlobanov/ogu/ogu.cgi	2004	FoldUnfold is based on the idea that the structure of proteins is governed by the balance between the interaction energy of residues and their conformational entropy.
IUPred [50, 51] http://iupred.enzim.hu	2005	IUPred is based on the idea that inter-residue interactions are responsible for determining whether a protein forms structure or not.
RONN [55] http://www.strubi.ox.ac.uk/RONN	2005	RONN is based on the functional alignments.
DISpro [56] http://scratch.proteomics.ics.uci.edu/	2005	DISpro, using a one dimensional recursive neural network (1D-RNN) model, combines the flexibility of Bayesian model with a fast, convenient, parameterization of an ANN.
FoldIndex [42] http://bip.weizmann.ac.il/fldbin/findex	2005	FoldIndex is used to analyze the ratio of net charge with hydropathy locally.
Spritz [59] http://distill.ucd.ie/spritz/	2006	Spritz consists of two specialized binary classifiers, one for short disordered regions and the other for long disordered fragments.
DisPSSMP [57]	2006	DisPSSMP is based on Radial Basis Function Networks with inputs from position-specific scoring matrices and other sequence properties.
IUP [103]	2006	A Recursive Maximum Contrast Tree (RMCT) was used to recognize intrinsically disordered regions.
DisPSSMP2 [58]	2007	DisPSSMP2 uses a two-level prediction scheme and a condensed position-specific scoring matrix.
PrDOS [60] http://prdos.hgc.jp/cgi-bin/top.cgi	2007	PrDOS consists of two predictors, one of which uses the alignment of homologs.
NORSnet [104]	2007	NORSnet uses feed-forward neural networks

Table 1 Predictors of IDPs

Predictor	Publication year	Brief description
		trained in a set of long loop regions.
POODLE-S [101] http://mbs.cbrc.jp/poodle/poodle-s.html	2007	POODEL-S is a group of seven SVM predictors with each responsible for a specific region of the whole sequence.
POODLE-L [102] http://mbs.cbrc.jp/poodle/poodle-l.html	2007	POODLE-L is composed of ten two-level SVM predictors.
POODLE-W [105] http://mbs.cbrc.jp/poodle/poodle-w.html	2007	POODLE-W predicts disordered structures by using a Spectral Graph Transducer (SGT) and by training with a huge amount of structure-unknown sequences.
Bayes [106]	2008	Bayesian method computes the conditional probability of a sequence from a certain class and then infers the posterior probability of the class
OnD-CRFs [107] http://babel.ucmp.umu.se/ond-crf/	2008	Conditional Random Fields (CRFs) method predicts the intrinsic disorder in proteins. CRF is a discriminatively supervised machine-learning method.
DISOclust [111] http://www.reading.ac.uk/bioinf/DISOclust/DISOclust_form.html	2008	DISOclust applies the principle that ordered residues within a protein target should be conserved in three-dimensional space within multiple models, whereas the residues that vary or are consistently missing may be correlated with the disordered structure.
metaPrDOS [109]	2008	MetaPrDOS is composed of seven individual predictors which areas follows: PrDOS, DISOPRED2, DisEMBL, DISPROT, DISpro, IUPred, and POODLE-S.
MD [110] http://cubic.bioc.columbia.edu/newwebsite/services/md/index.php	2009	MD is a metapredictor composed of NORSnet, Ucon, PROFBval, DISOPRED2, IUPred, and FoldIndex.
CDF-ALL [108]	2009	CDF-ALL is a protein-level disorder predictor composed of CDFs from VLXT, VSL2, VL3, TopIDP, IUPred, and FoldIndex.

can be divided into three periods.

The first period (1) First informal IDP predictor The first period is the beginning of the prediction for IDPs and includes the predictors designed before 2002. In 1979, much earlier than the publication of the PONDR®s, Williams [34] made the first attempt to predict lack of structure based on amino acid sequence. He suggested that the ratio given by (number of charged amino acids)/(number of hydrophobic amino acids) would distinguish non-folding proteins from structured examples. This suggestion was based on very good results for a small number of examples. However, Williams did not follow up his approach with a larger number of proteins so as to make a quantitative estimate of the accuracy of his predictor. Therefore, we tested the Williams ratio as a disorder predictor, using the charged and hydrophobic amino acids suggested by Williams and using sets of fully structured and fully disordered proteins containing hundreds of examples in each class. To our disappointment, the C/H ratio turns out to be a very poor predictor of disorder. A large fraction of structured and disordered proteins have overlapping values for the Williams ratio (Bin Xue, unpublished observations).

(2) Development of the PONDR® family The first

formal predictor was published by Romero *et al.* [36] in 1997, which could predict disordered structure in proteins based on the specific biases of their amino acid sequences. These predictors were subsequently called Predictors of Natural Disordered Regions (PONDR®) followed by letters that describe the proteins in the training set for that particular predictor.

Basically, the PONDR® algorithms work because the amino acid compositions in a window of N amino acids for structured proteins are distinguishable from the compositions for disordered proteins. A typical PONDR® uses the following types of inputs: (1) amino acid compositions; (2) attributes derived from compositions such as sequence complexity; and (3) attributes derived from compositions via some function or scale such as hydropathy, net charge, etc. These various types of attributes are then weighted and combined in a non-linear manner, typically via artificial neural networks (ANNs). Other methods of combining the attributes such as support vector machines (SVMs) and logistic regression give results very similar to those obtained with ANNs.

Obradovic and coworkers used three different algorithms, logistic regression, discriminant analysis, and ANN, to predict disordered structures of proteins [37]. ANN gives a slightly higher accuracy. But prediction accuracy is only a simplistic indicator, and it is inappropriate to rank the methods on this basis alone. Logistic regression represents the most robust method for predicting two states, order and disorder.

One of the earlier examples is PONDR® VL-XT, where VL describes a training set of "Variously characterized Long" (> 30 residues) disordered regions, and two additional training sets of X-ray-characterized Terminal regions, one for the amino-terminus and one for the carboxy-terminus [38]. This division is based on a hypothesis that the disordered structure characteristics of sequences might depend on the location of the disordered region in the sequence.

VL3-E is a combination of two ANN-based predictors, VL3-H and VL3-P. VL3-H searches for homologous sequences to increase the number of examples in the training sets, while in VL3-P profiles of a sequence generated by PSI-BLAST are added as an input attribute to improve the accuracy of predicting disordered regions [39].

The most recent advance in the PONDR® collection is the set of VSL predictors (trained on Variously characterized, Short and Long disordered regions). The very first publication from the PONDR® developers pointed out that short and long disordered regions might have differences in their amino acid characteristics because predictors trained on short regions of disorder did poorly on long regions of disorder and vice versa [89]. The VSL predictors take advantage of this difference. Two versions of PONDR® VSL have been developed. In the development of VSL1, both neural networks and ensembles of logistic regression models were tried, with one neural network or one ensemble trained on short (≤ 30 residues) regions of disorder and a second network or a second ensemble trained on long (> 30 residues) regions. Irrespective of whether neural networks or logistic regression ensembles were used for the short- and long-region predictors, the two predictors were combined by means of a logistic regression model. Since the neural networks and logistic regression ensembles gave similar prediction accuracies, only the simpler, logistic repression models appeared in the subsequent publication [40]. In the VSL2 predictor, the short- and long-region predictors were replaced with SVMs, but likewise a logistic repression model was used for the merger [41].

VSL1 does reasonably well for predictions of both long and short disordered regions. It not only achieves the same or higher accuracy as most other predictors for long disordered regions, but also improves significantly the prediction accuracy of short disordered regions. Like VSL1, VSL2 also has the purpose of addressing the length-dependency problem in disorder prediction. Its results further confirm the differences in amino acid compositions and sequence properties between short and long disordered regions. Short disordered regions are more depleted in I, V, and L, while long disordered regions are more enriched in K, E, and P but are less enriched in Q. In addition, long disordered regions are depleted in G and N, while short disordered regions are enriched in G and D. Figure 4 shows the difference in amino acid compositions between short and long disordered regions [41].

The VSL1 predictor was evaluated as the highest ranked overall among those predictors at CASP6 [84] and VSL2 was evaluated as the highest ranked overall in CASP7 [85], both in terms of prediction accuracy and in terms of the area under the ROC curve. As stated above, given the small number of test examples in the CASP exercise, no claim is being made that VSL2 is currently the best disorder predictor. These predictors are discussed in more detail below.

(3) CH-plot In 2000 Uversky *et al.* used charge and hydropathy to predict disorder, but in an entirely new way compared to approach used by Williams much earlier as described above. The Uversky *et al.* [35] approach is based on the simple reasoning that the folding of a protein is governed by a balance between attractive forces (*e.g.*, hydrophobic interactions) and repulsive forces (Coulomb or electrostatic repulsion). Rather than the total charge used in the Williams ratio, net charge

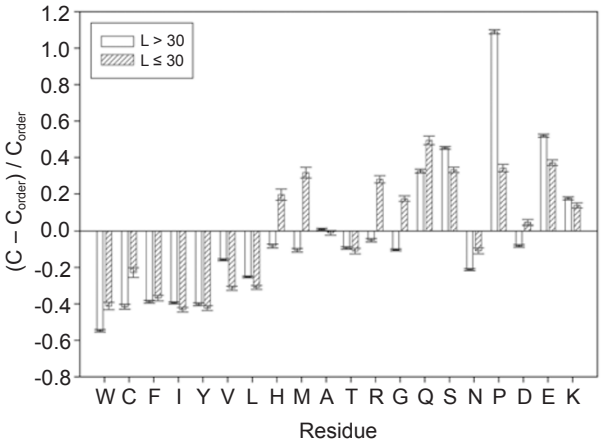

Figure 4 Relative amino acid compositions for short and long disordered regions. The set of globular proteins, globular-three-dimensional, the set of short (30 residues or shorter) and the set of long intrinsically disordered regions (longer than 30 residues) in proteins from the DisProt database (version of 17 October 2008) are compared. The fractional difference was calculated as $(C - C_{Corder})/C_{order}$, where C is the content of a given amino acid in a given protein set and C_{order} is the corresponding content in a set of ordered proteins, and plotted for each amino acid. In this plot, the amino acids are arranged from the most order-promoting to the most disorder-promoting. Enrichment and depletion in each amino acid type appears as a positive and negative bar, respectively. Amino acids are indicated by their single-letter code. Confidence intervals were estimated using per-protein bootstrapping with 1 000 iterations.

is more appropriate for estimating repulsion. High net charge leads to strong electrostatic repulsion and low hydropathy minimally means less driving force for compaction. Thus, whether a query sequence is disordered or not can be predicted based on the ratio of its mean net charge with mean hydropathy. The mean hydropathy is defined as the sum of the hydropathies of all residues divided by the number of residues in the polypeptide. The mean net charge is similarly defined as the net charge, at pH 7.0, divided by the total number of residues. A plot of mean net charge versus mean hydropathy (the CH-plot) separates structured and disordered proteins into distinct regions and thereby provides a binary predictor [35]. Prilusky *et al*. [42] used this idea to design a per-residue disorder predictor, FoldIndex. By computing the CH ratio along the protein, this predictor can predict if a local region in given sequence is in a disordered structure.

The second period The second period covers 5 years from 2002 to 2006. Many predictors were developed during this period, including VL3-E, and VLS1 and VLS2 of PONDR®. Two other early predictors are GlobPlot and DisEMBL.

The kernel idea of GlobPlot is the relative propensity of an amino acid residue to be in an ordered or disordered state [43]. GlobPlot uses an amino acid scale based on the difference in the probability for a given amino acid to be in random coil or to be in regular secondary structure. The basic algorithm behind GlobPlot is simple and very fast, representing a sum function. In order to smooth the curve of this function, a digital low-pass filter based on the Savitzky-Golay algorithm is run. Then the numerical estimation of the first order derivative is retrieved. The resulting smoothed function is plotted using the DISLIN 8.0 package. Putative globular and disordered segments can be selected by using a simple peak finder algorithm. Generally speaking, the change of slope corresponds to the boundary between the ordered and disordered structures. Therefore, this method can also be utilized to identify globular domains.

DisEMBL designed by Linding *et al*. [43] consists of three separate ANN predictors, to predict three kinds of disordered structures in proteins, which represent residues within "loops/coils (as defined by DSSP38)", "hot loops (loops with high B-factors)", or those that are missing from the PDB X-ray structures (called "Remark 465"). Linding *et al*. also investigate the relationships between the different disorder definitions. The prediction results indicate that the hot loops show less correlation with coils and more with the Remark 465 examples. Much more work still needs to be done for a deeper understanding of relationships among the various disordered structures, which could lead to an improved definition of IDPs.

An important event was the inclusion of IDP prediction as a category at CASP5. CASP has subsequently played a very positive role in promoting the development of IDP prediction. During this period, CASP has held a total of three exercises involving disorder prediction, and each event was accompanied by the emergence of several new IDP predictors. First, we introduce DISOPRED designed by Jones *et al*. [45]; this predictor achieved a Wilcoxon score of 90.0 at CASP5.

DISOPRED employs a feed-forward neural network to predict disordered regions [45]. It needs to use a relatively large number of hidden units to enhance the mapping ability, and this may cause over-fitting and slow the training down. So, they designed DISOPRED2 based on SVMs instead of ANN [47]. SVMs can improve generalization by controlling the classifier's capacity and the associated potential for over-fitting [47]. Compared with other methods for disordered structure prediction, the main difference of DISOPRED2 is that it directly trains on the whole sequence rather than measures of amino acid composition, sequence complexity, or biophysical

properties such as mean hydrophobicity. The training set for DISOPRED2 contained 715 ordered proteins with high-resolution X-ray structures (better than 2.0 Å) and less than 25% pairwise sequence identity. This set included 176 550 ordered and 4 590 disordered residues. For each protein in the training set, a sequence profile was generated using three iterations of a PSI-BLAST search against a non-redundant sequence database. The predictor was trained using various combinations of binary-encoded amino acid sequence, secondary structure predictions (SSPs) from PSIPRED, and PSI-BLAST profiles for symmetric windows of 15 positions. The N- and C-termini were treated separately.

Subsequently, Jones and coworkers used DISOPRED2 to estimate the frequency of disordered structures in several representative genomes from the three kingdoms of life [46]. The prediction result obtained in this study is consistent with earlier studies, indicating that IDPs very commonly exist in eukaryotes but less so in prokaryotes [90]. To explain this result, several explanations have been proposed. Prokaryotes are subject to strong selective pressure on biochemical efficiency and do not have highly regulated degradation pathways such as ubiquitination, so the cost of short protein lifetimes is likely to be far greater. The absence of cell compartments may also reduce the ability of prokaryotic cells to physically protect disordered structures from degradation.

Perhaps due to the promotion of CASP, there has been recently increased interest in predicting disorder and associated features from sequence. One example is that of Weathers *et al*. [91] who use reduced sets of amino acids to predict IDPs; this approach attained a high accuracy. The success in the prediction of IDPs has indicated that the composition of amino acid sequences may be a reliable feature to indicate the presence of disorder. Weathers *et al*. [91] demonstrate that not only amino acid composition but also reduced sets of amino acids based on chemical similarity were able to achieve a high accuracy for the IDP prediction. This approach aims to predict disorder simply as a linear combination of the composition vectors using either the full or a reduced amino acid alphabet. In more detail, each protein in the dataset of 1 190 ordered proteins and 718 disordered segments was translated into a vector representation. For the full amino acid alphabet, the vector set was based on sequence composition information for each amino acid. Therefore, proteins were represented with one vector for each amino acid, thereby leading to a 20-amino acid SVM. For reduced amino acid alphabets, proteins were described by sets of 15, 10, 8, and 4 vectors. In this analysis, additional SVMs were developed to find optimal weights by taking linear combinations of composition vectors, *e.g*., a dot kernel function, $K(s_i, x)=s_i \cdot x$, was used in the process to map the sample data into a higher-dimensional space, where s_i is a support vector and x is the input sequence. The reason for selecting this kernel function was that it not only provides a high accuracy but also avoids the long training and testing time associated with the higher order kernel functions. The prediction results demonstrated that the reduced composition of amino acids can also gain high prediction accuracy. Even the reduced set as small as four maintained a high prediction accuracy. Thus, the composition of amino acid sequences can be considered as an important factor contributing to the disorder prediction of proteins.

During the years from CASP6 to CASP7, the number of published predictors grew more than that at any other time. Some of them have been mentioned above, including VSL1 and VSL2 that have been evaluated as the best at CASP6 and CASP7, respectively. Others, such as DRIPPRED [49], IUPred [50, 51], FoldUnfold [52-54], RONN [55], DISpro [56], DisPSSMP [57, 58], and Spritz [59], were also proposed in these years.

MacCallum *et al*. [49] designed DRIPPRED based on Kohonen's self-organizing map (SOM), which was generated for a non-redundant set of UniProt sequence profiles. First of all, selected data were made non-redundant using a crude single-pass Perl hashing approach. Every sequence of length L residues in the generated protein data was run through PSI-BLAST in order to obtain the position-specific scoring matrix (PSSM). Then profile windows of sequences were mapped into an SOM. Through the training process, every sequence profile window can be mapped into a discrete position in the SOM grids. Predictions were based on hit frequencies to a certain area in this map. Sequences that mapped to part of the UniProt space that were relatively unpopulated by proteins of known structure were assumed to correspond to disordered regions. After that, the frequency of different types of amino acids in the SOM nodes was calculated and the classifiers were designed, which can determine the prediction result of the query protein.

Dosztanyi *et al*. suggested that a large number of inter-residue interactions is responsible for structure stabilization of proteins [50, 51]. In contrast, IDPs don't have sufficient numbers of stabilizing inter-residue interactions. Based on this reasoning, an IUPred algorithm estimating the inter-residue interactions was designed. First, the interaction energy between each pair of amino acids based on their C_β positions was estimated. This was done by calculating the potential mutual contact energies for all amino acid pairs in a dataset of globular proteins with known structure. This is a fairly standard approach in computational biology, and in this work Dosztanyi *et al*.

compared several such mutual contact energies estimated previously by other researchers, with the set developed by Thomas and Dill, found to be the best in this particular application [92]. The various pairwise energies were assembled into a 20 × 20 energy matrix, which was used in the next step, the estimation of the mutual interaction energies for any given protein. The prediction utilizes this energy prediction matrix and the amino acid compositions put into a quadratic expression. These statistical values represent the ability to form stabilization contacts between amino acids in polypeptide chains. The potential mutual interactions were estimated using amino acid compositions, not three-dimensional structures. These composition-based energies were compared with three-dimensional structure-based energies of the proteins for which the actual side chain interactions are known. The composition-based potential mutual interaction energies and the structure-based energies were found to be highly correlated, thus the former can be used to estimate the latter even when the structures are not known. To use this approach to predict structure or disorder, composition-based calculations for a set of proteins that fold into three-dimensional structures were compared with composition-based calculations for a set of disordered proteins. The estimated potential interaction energies for the structured proteins were much greater than the same energies for the unstructured proteins, and from these results the energy boundary between ordered and disordered proteins as a function of length was determined. This boundary allows the recognition of intrinsic disorder. In brief, if a sequence contains too few hydrophobic residues, then the composition-based potential mutual interaction energy will necessarily be too small and thereby indicate the lack of potential for folding.

Galzitskaya *et al*. think that the formation of ordered structure in proteins is mainly determined by the balance between the interaction energy of residues and their conformational entropy [52-54]. They designed the FoldUnfold predictor based on this idea. In this work, an interesting parameter, namely the mean packing density of residues, is used to express the average contact number of residues within a given distance in a protein structure. It has been demonstrated that regions with low-expected packing density correspond to the disordered fragments. Interestingly, residue packing density values (called residue contact values at that time) were used in early disorder predictors developed by the PONDR® group [38]. The use of this feature was based on a scale that appeared in a technical report [93], not in a published paper. In feature selection experiments, this scale was found to be very promising, but its use was discontinued in later versions of PONDR® due to our failure to find a published and readily available version of this scale at that time. The recent success of FoldUnfold suggests that it would be useful to investigate the reintroduction of this feature into future PONDR®s.

Yang *et al*. [55] designed RONN to predict disordered structures based on the sequence alignments. In general, it is assumed that similar sequences are likely to have similar functions (*e.g*., being ordered or disordered). Therefore, a similarity to prototype disordered sequences is evaluated and predictions are made based on a function of these values. Suppose the sequences of a group of ordered and disordered proteins are known, the disordered status of a query sequence can be inferred by comparing it with all the known sequences. In the training process, the similarity of sequences is evaluated by sequence alignment techniques using a mutation matrix to score the similarity. These scores of sequence alignments are then used for training. After training, every sequence can be classified as being ordered or disordered. In the testing process, if the testing result can satisfy the desired accuracy level, the modeling then progresses to the prediction process. Alternatively, the training process is repeated until the accuracy reaches the pre-defined level. In the prediction process, the sub-sequence of the query sequence needs to be aligned with all prototype sequences to get the homology scores. Using the above model, a probability of disordered structure formation for every query sequence can be evaluated. RONN is imperfect in that it doesn't do well in predicting short disordered regions, nor the first and last residues of disordered regions. In fact, most predictors for disordered structures have this problem. The prediction of disordered structures in these regions has become one of the important issues now.

DISpro, using a 1D-RNN model, combines the flexibility of a Bayes model with the fast, convenient, parameterization of ANN without the shortcoming of the standard ANN feedback with fixed input size [56]. The input includes 25 attributes, 20 corresponding to the amino acid frequencies, 3 corresponding to the predicted secondary structure class of the residue, and the last 2 corresponding to the predicted relative solvent accessibility of the residue. Because the older version of DISpro needed a pre-defined threshold to classify output results, it was not able to show the relationship between sensitivity and specificity. Lately, this predictor has been improved so that the threshold can be changed at will [94]. In this way, users can investigate the relationship of specificity and sensitivity for disordered and ordered residues.

Su *et al*. [57] studied the effect on the prediction of disorder by using a condensed PSSM obtained from PSI-BLAST with respect to the physicochemical properties

(PSSMP). Based on this study DisPSSMP was designed to predict the disordered structures of proteins. A derivative of this predictor, DisPSSMP2, is a two-stage classifier that further enhances the prediction power of DisPSSMP [58]. Both of these predictors employ Radial Basis Function Networks, the output of which denotes a probability that a given residue is in a disordered state. In DisPSSMP, if the output result of a residue is higher than a given threshold, this residue is predicted to be in a disordered state, whereas in the DisPSSMP2, these output results are collected to compose a set for the second layer prediction that redefines and adjusts the threshold value and size of sliding window, thus smoothing the results of the first layer.

Spritz is implemented by using two specialized binary classifiers, one for short disordered regions and the other for long disordered fragments [59]. The purpose of doing so is to develop different, disjoint expertise by taking advantage of the different class distributions in the two cases. Spritz uses an SVM with a non-linear kernel function. The frequencies of twenty amino acids are used as the inputs into the SVM. The Spritz server has two interfaces, one for single and one for multiple queries.

As many viral proteins have a modular organization, containing regions (hydrophobic or disordered) that are often not compatible with the crystallization process [95, 96], the 'viral enzyme module localization' (VaZyMolO) tool to define and classify viral protein modularity was elaborated [97]. VaZyMolO analyzes viral proteins by implementing BLAST [98], multalin [99], hydrophobic cluster analysis (HCA) [100], and CH-plot analysis [35]. VaZyMolO is organized to have three layers reflecting surface (layer S), matrix (layer M) and non-structural proteins (layer F), and aims at defining viral protein modules that might be expressed in a soluble and functionally active form, thereby identifying candidates for crystallization studies [97].

POODEL-S [101] is another complex of predictors. This predictor is composed of seven SVM predictors, with each responsible for a pre-determined region away from both termini. Each SVM is trained by various combinations of 10 physicochemical features and sequence profiles from PSI-BLAST. For amino acids overlapped by different regions, the average of the predictions from each region is taken as the final prediction value. POODLE-L [102] is a similar two-level SVM predictor that divides the whole sequence into 10 subregions. The inputs of this predictor are only 10 physicochemical properties. The first level predicts the disordered probability for a segment, and based on the output of the first level and the physicochemical properties of each amino acid, the second level presents a residue-based prediction.

A method based on the Recursive Maximum Contrast Tree (RMCT) was also used to recognize IDPs [103]. This classifier utilizes K-Nearest Neighbor Decision Rules, where the nearest neighbors are defined by the tree structure. Classifications with RMCT on tree nodes are guided by K majority voting principles. First, using a particular feature to be tested, this algorithm calculates the distances between a test instance and two sets of training instances, and then it calculates the average distances for the two sets of test instances. The two sets of training instances are comprised of ordered and disordered regions. The overall distance is then calculated as the difference between the average distances obtained for the ordered and disordered sets divided by the square root of the sum of the standard deviations calculated for the distributions of the distances from the test instance to the training instances. A predictor is then formed by a majority vote over a set of features selected to form a given predictor. By combining the decisions from many different predictors, the overall performance was further improved [103].

NORSnet [104] is another feed-forward neural network-based predictor for long disordered loop regions. The inputs of this method include the sequence profile from PSI-BLAST and a group of predictions of secondary structure, solvent accessibility, and flexibility, as well as various attributes related to the sequence composition. The largest difference of this method from other disorder predictors is that it is trained on a specially selected set of long loop regions.

The third period In the previous two periods, the prediction techniques mainly included ANNs and SVMs. IDP prediction after CASP7 exploited more methods, such as Spectral Graph Transducers (SGTs) [105], Bayesian methods [106], Conditional Random Fields (CRFs) [107], and metapredictors [61, 108-110].

Shimizu *et al*. [105] proposed the use of information from structure-unknown proteins in order to avoid training data sparseness. They predicted disordered structures by using a SGT, with training on a huge amount of structure-unknown sequences as well as structure-known sequences. The SGT was used to construct a k-nearest-neighbor graph, which takes into account the information on the unlabeled data. SGT assigns a label to U by dividing G into two subgraphs, G^+ and G^- . The prediction results show that the data with structure-unknown information can not only expand the training set but also improve the accuracy of the disorder prediction.

The prediction system of PrDOS consists of two predictors, one based on the local amino acid sequence information and the other based on the template proteins [60]. First of all, the target amino acid sequence is con-

verted into a PSSM. Then, two predictions are performed using the PSSM. The first SVM-based predictor is built based on the local amino acid sequence information. The second predictor is based on template proteins and uses the alignments of a query sequence with structures that are known. To combine the results of two independent predictors, the weighted average between the results of two predictors is calculated. PrDOS uses the alignment of homologs with templates that have been determined. The alignment of homologs is very popularly applied to predict the secondary structures of proteins. The success of PrDOS indicates that it may provide a useful reference for the disorder prediction.

In view of the above concepts provided by Dosztanyi *et al*. and Galzitskaya *et al*. that average contact propensities can be used to predict disordered structures, Schlessinger *et al*. [104] make use of protein-specific internal contacts to predict disordered regions relevant for protein interactions and thereby designed Ucon. Ucon is a specific predictor that only has the ability to identify long disordered regions (> 30 residues in length).

Bayesian classifier methods have wide applications in the structure prediction of proteins. Bulashevska and Eils [106] were the first to apply this approach to predict IDPs. Each protein sequence belonging to a certain class can be considered as a realization of an independent random process that emits symbols from an alphabet of 20 amino acids. In this classifier, the appearance of every amino acid is considered as an independent event. This method computes the conditional probability of a sequence from a certain class, and then infers the posterior probability of the class for an unlabeled sequence based on Bayes' rule. In this analysis, the attributes used include the compositions of the amino acid sequences. Since the amino acid composition depends on the length of disordered regions, they make three separate representations to predict long, medium, and short disordered regions. This predictor achieves good performance.

Wang *et al*. [107] used a new method, CRFs, to predict the intrinsic disorder in proteins. A CRF is a discriminatively supervised machine-learning method. Compared to ANNs and SVMs, CRFs are able to take into account interrelation information between two labels of neighboring residues. The features of amino acid sequences and information on predicted secondary structure are used as the inputs of this model. A limitation of this approach is that the training speed of CRFs is slow.

The DISOclust method is based on the simple premise that the ordered residues within a protein target should be conserved in three-dimensional space within multiple models, whereas the residues that vary or are consistently missing may be correlated with the disordered structure [111]. This method can be divided into two steps, the prediction of the per-residue error in multiple fold recognition models and a simple analysis of the conservation of per-residue error across all models. At the first step, the per-residue quality of each model is calculated by carrying out structural alignments with every other model using the TM-score program. The average S-score of each kind of residue in each model is calculated. These scores of each model are added together and divided by the number of models. Then a mean S-score for each residue in all models can be evaluated. The approximate posterior probability of a residue being in a disordered state can be expressed as 1 minus this score. McGuffin demonstrates that a simple consensus of methods that includes DISOclust can significantly outperform all of the previous individual methods tested.

Recently, a new direction in the development of disorder predictors based on the creation of metapredictors has attracted attention. These metapredictors combine the outputs of several individual predictors. They can be applied either at the residue level or at the whole sequence level. Often, the individual predictors constituting metapredictors use different philosophies for prediction. In the following, we discuss one older metapredictor followed by discussion of three very new metapredictors developed during the most recent period of predictor development.

For the first time, two philosophically different predictors were combined in a metapredictor in 2005 [61], when a consensus method was developed that was based on two distinct binary classifiers, the CH-plot [35] and the CDF analysis [90]. Carefully selected sets of 52 wholly disordered and 105 unique, wholly ordered monomers without ligands or disulfide bonds were used in this study. Furthermore, a set of 64 partially ordered proteins was derived from PDB structures that contained a single chain and a unit cell with a primitive space group. As mentioned above, the CH-plot discriminates ordered and mostly disordered proteins based on the combination of net charge and hydrophobicity [35]. A simultaneous observation of low mean hydropathy and relatively high net charge is typical for the "natively unfolded" proteins, which are characterized by the lack of compact, collapsed structure. Therefore, ordered and disordered proteins plotted in CH-space can be separated to a significant degree by a linear boundary, with proteins located above this boundary line being natively unfolded and with proteins below the boundary line being ordered [35]. The CDF analysis was proposed as a method for classifying proteins as being mostly ordered or mostly disordered based on the per-residue PONDR® VL-XT outputs [90]. The CDF summarizes the per-residue disor-

der predictions by plotting PONDR® scores against their cumulative frequency, which allows ordered and disordered proteins to be distinguished based on the distribution of prediction scores. In more detail, the CDF curve gives the fraction of the outputs that are less than or equal to a given value. According to the CDF analysis, fully disordered proteins have low percentages of residues with low-predicted disorder scores, since the majority of their residues possess high-predicted disorder scores. On the contrary, the majority of residues in ordered proteins are predicted to have low disorder scores. Hence, theoretically, the curves for all the fully disordered proteins should stay at the lower right quadrant of the CDF-plot, whereas all the fully ordered proteins should be located at the upper left quadrant [88, 90]. Therefore, overall, the CH-plot is a linear classifier that takes into account only two parameters of the particular sequence – charge and hydropathy, whereas CDF analysis is dependent upon the output of the PONDR® VLXT predictor, a non-linear neural network classifier, which was trained to distinguish order and disorder based on a significantly larger feature space that explicitly includes net charge and hydropathy. According to these methodological differences, CH-plot analysis is predisposed to discriminate proteins with substantial amounts of extended disorder (random coils and pre-molten globules) from proteins with globular conformations (molten globule-like and rigid well-structured proteins). On the other hand, PONDR®-based CDF analysis may discriminate all disordered conformations including molten globules from rigid well-folded proteins. Therefore, this discrepancy in the disorder prediction by CDF and CH-plot might provide a computational tool to discriminate "natively unfolded" proteins from native molten globules, which might be predicted to be disordered by CDF, but compact by CH-plot [23, 61]. This model is consistent with the behavior of several IDPs. Next, the CH- and CDF-plots were combined into a single classification method using the consensus scoring method that focuses on correct classification of proteins for which prediction methods disagree by using a weighted combination of the reliability measures [61].

The work described above was very recently followed up, and this led to an improved binary metapredictor to estimate whole protein structure or disorder [108]. This new metapredictor was based on a combination of several CDF predictors developed from several disorder predictors, including PONDR®s VLXT, VSL2, and VL3, TopIDP, IUPred, and FoldIndex. A neural network was then used to combine these individual CDF-based predictions. The neural network was trained on a fully disordered subset obtained from DisProt and a fully ordered dataset extracted from PDB. In comparison with the individual whole protein predictors, this metapredictor improved the prediction accuracy by 5%-10% on various datasets [108].

Ishita and Kinoshita developed the metaPdDOS metapredictor for per-residue estimates of order and disorder [106]. The MetaPdDOS uses a SVM to integrate residue-level predictions from PrDOS, DISOPRED2, DisEMBL, DISPROT, DISpro, IUPred, and POODLE-S. This SVM was trained on a group of PDB-extracted proteins that all have regions of missing electron density in their crystal structures, and the sequence identities among these proteins are less than 20%. By using only two components, PrDOS and DISpro, this metapredictor achieves an accuracy in terms of the AUC (area under the ROC curve) of about 0.897 estimated by 10-fold cross-validation. By utilizing all seven individual predictors, the 10-fold cross-validation AUC goes up to 0.904 ± 0.004. The AUC of this method on the CASP7 dataset is also high, giving 0.877 ± 0.007. These researchers did not identify a specific threshold for their metapredictor, and so did not report an overall accuracy for order-disorder prediction.

Based on the important observation that the reliability of disorder prediction benefits from the use of several methods relying on different concepts or different physicochemical parameters [79, 81, 97], a web metaserver MeDor for fast, simultaneous analysis of a query sequence by multiple predictors was developed [112]. MeDor provides a graphical interface with a unified view of the outputs of the following programs: a SSP, based on the StrBioLib library of the Pred2ary program [113, 114], HCA [100], IUPred [51], Prelink [115], RONN [55], FoldUnfold [53], DisEMBL [44], FoldIndex [42], GlobPlot2 [43], PONDR®s VL3 and VL3H [116], PONDR® VSL2B [40], and Phobius [117]. The authors emphasize that MeDor does not provide a consensus of disorder prediction. The major goal of this web metaserver is to provide a global overview of various predictions utilizing different philosophies, and to accelerate the process of disorder prediction by multiple tools [112].

Schlessinger *et al*. [110] designed another metapredictor named MD (MetaDisorder) predictor. This method employs neural networks to combine the predictions from NORSnet, DISOPRED2, PROFbval, and Ucon. A second method described in this work also combines the prediction from FoldIndex, IUPred, and several additional sequence features, such as predicted secondary structure, local sequence profile, predicted solvent accessibility, sequence complexity, amino acid composition, sequence length, etc. The training datasets were proteins from PDB and DisProt. This method achieved an AUC of 0.80, which is several percentage points higher than

the values estimated for the individual predictors. Before comparing the performance of this metapredictor with those discussed above, one should remember that comparing predictors is typically equivocal due to differences in the training and testing datasets, and due to variations in the methods used for accuracy evaluation.

Difficulties of IDP Prediction

The enthusiasm for predicting disordered structures of proteins continues to grow. As mentioned above, the number of representative predictors has increased to more than 40 since the first one was published in 1997. In the process of designing and using these predictors, much significant biological and biomedical information has been obtained. The related findings have been analyzed in detail within the descriptions of the corresponding predictors. Although the IDP-related research is continuing to be of significant interest, and, although gratifying achievements occurred, this field is still experiencing many difficulties.

First, the number of proteins with experimentally determined disordered structure is still small. A databank of IDPs has been built [33]. The DisProt Release 2.0 (14 February 2005) included 179 IDPs and 290 disordered regions, whereas the DisProt Release 4.5 (7 July 2008) included 520 IDPs and 1 191 disordered regions. This indicates that the number of annotated IDPs is increasing rapidly, but there still is an enormous gap compared with the actual number of IDPs in nature and compared to the number of experimentally validated IDPs. The prediction results of PONDR® and DISOPRED2 have revealed that IDPs are very common in eukaryotes. The number of IDPs experimentally characterized so far is only a tiny fraction of this total, and of the experimentally characterized examples, only a small fraction of these have been annotated within the current database. The lack of sufficient numbers of structurally characterized IDPs significantly limits the development of new predictors and also limits the ability to improve already existing predictor algorithms. Clearly, an increased level of resources devoted to the annotation of IDPs is important.

Second, because the molecular mechanisms of disordered structure formation are not very clear yet, the selected characteristics and attributes in the prediction process are mainly focused on the biochemical properties of amino acids and on the compositional content of sequences. These current approaches might be putting limits on the descriptions of structural peculiarities of IDPs, and this in turn might cause us to underestimate the frequency of IDPs. These observations suggest that a large increase in the amount of experimental work on these proteins should be done to provide more biomedical and biophysical information regarding IDPs.

Third, IDPs can be characterized by more than 20 biophysical methods, each of which gives slightly different information. Thus, if time and money were no issue, there would be a significant advantage to characterize a significant number of disordered regions by multiple methods [110, 118]. A collection of IDPs characterized by multiple methods would provide an important basis set for developing a deeper understanding of different types of disorder.

Fourth, a consensus has not yet been reached among different researchers regarding even the definition of IDPs. Typically researchers involved in the IDP studies have put forward different IDP definitions based on different aspects of the protein structural ensembles, and several researchers have pointed out the importance of characterizing individual IDPs by a variety of computational methods [5, 61, 79, 81, 97, 110, 112, 118]. But even if this were done, the situation would still resemble the Indian parable of collection of blind men examining an elephant, with each method providing a different impression of the shape and structure of the beast. A very important development will be to learn how to merge the various types of information regarding disordered proteins into a common model or into a common set of models. In this regard, the methods to identify or classify different types of disorder, especially different types that are associated with different functions [119], need to be improved.

Future of IDP Prediction

Accurate IDP prediction is a necessary prerequisite for the complete understanding of the principles of protein folding. IDP prediction is also needed for comprehension of the molecular mechanisms of protein function and for building a new structural and functional hierarchy of proteins. Recently, researchers have found that IDPs often have a close relation with some diseases including cancer, cardiovascular disease, diabetes, neurodegenerative disease, and amyloidoses [120]. Thus, accurate IDP prediction is drawing more and more attention. At present, the developing directions of IDP prediction include three aspects as follows:

(1) Improve the accuracy of the prediction for IDPs, especially short disordered regions and disorder in N- and C-terminal regions of protein sequences. In the prediction process, these regions are often ignored by researchers. As a result, a large number of rather accurate algorithms are currently designed to predict long disordered regions, whereas the prediction accuracy for

the short and terminal disordered regions remains lower. Overall, the prediction of short and terminal regions is challenging and clearly represents one of the crucial directions for future development.

(2) Another difficult problem that prevents improved prediction for IDPs is the high noise level regarding both the structured and disordered regions that are used as training sets. Structured proteins often have localized regions that undergo disorder-to-order transition upon complex formation or even upon crystallization. Likewise, disordered proteins have local subregions that are primed to form structure when interacting with a partner, and so, from the sequence point of view, these regions often have the characteristics of structured proteins. Thus, experimentally characterized regions of structure and disorder both have significant levels of noise with regard to the correct assignment of order or disorder.

(3) Currently, in the IDP prediction field, the most commonly used computational techniques are ANN and SVM, which could be considered as "black-boxes." The prediction results of these "black-box" models are not easily understood in terms of their underlying sequence features. On the other hand, methods based on physicochemical properties, such as CH-plot, IUPred, and FoldIndex, are still not so accurate. One approach that has been used to better understand the ability of a particular sequence to fold into a particular structure has been the application of protein-design approaches. To our knowledge, no one has carried out protein-design approaches to try to better understand the distinctions between disorder-forming sequences from those that form structure. Such approaches could potentially provide a new window into the reasons why sequences fold into three-dimensional structure or remain as poorly structured ensembles.

(4) Two general interrelated difficulties are likely to be important for future improvement of IDP prediction. These difficulties are as follows: (i) determining the most effective techniques for data classification and (ii) given a particular data classification, selecting the most appropriate approach for prediction. Yang *et al*. [55] suggested that incorporating some well-performing methods and looking for common disordered features may be the best way to gain the reliable identification of disordered regions of proteins. In essence, this approach is to classify IDPs into subsets based on coherent prediction by a well-performing method, and then to study the coherently predicted IDP subsets to discover their common features.

Summary Comments

Both equipment and labor costs are very high for laboratory-based experiments, and the results obtained from the use of a single type of experiment often give ambiguous analysis of IDPs. Given these limitations and difficulties in the experimental characterization of IDPs, prediction techniques are expected to continue to be very important for helping to develop an understanding of these proteins. With the continued development of prediction techniques, more accurate or at least better-understood IDP prediction is expected to occur. Given current or hopefully more accurate IDP predictors, an important future goal will be to improve sequence-function identification for IDPs and to include such IDP-based functional annotation in the amino acid sequences of the various model organism databases.

Acknowledgments

This work was supported in part by the grants R01 LM007688-01A1 (to AKD and VNU) and GM071714-01A2 (to AKD and VNU) from the National Institutes of Health and the Program of the Russian Academy of Sciences for the "Molecular and cellular biology" (to VNU). We gratefully acknowledge the support of the Indiana University Purdue University Indianapolis (IUPUI) Signature Center Initiative, developed by Executive Vice Chancellor Dr Uday Sukhatme and supported by funds from the Office of the Executive Vice Chancellor, IUPUI, Indianapolis, Indiana, USA. Dr Edwin Harper of the Department of Biochemistry and Molecular Biology, Indiana University School of Medicine, Indianapolis, IN, USA, is thanked for very helpful discussions.

References

1 Wu H. Studies on denaturation of proteins. XIII. A theory of denaturation. *Chin J Physiol* 1931; **1**:219-234.
2 Edsall JT. Hsien Wu and the first theory of protein denaturation (1931). *Adv Protein Chem* 1995; **46**:1-5.
3 Dunker AK, Oldfield CJ, Meng J, *et al*. The unfoldomics decade: an update on intrinsically disordered proteins. *BMC Genomics* 2008; **9**:S1.
4 Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradovic Z. Intrinsic disorder and protein function. *Biochemistry* 2002; **41**:6573-6582.
5 Radivojac P, Iakoucheva LM, Oldfield CJ, *et al*. Intrinsic disorder and functional proteomics. *Biophys J* 2007; **92**:1439-1456.
6 Vucetic S, Xie H, Iakoucheva LM, *et al*. Functional anthology of intrinsic disorder. 2. Cellular components, domains, technical terms, developmental processes, and coding sequence diversities correlated with long disordered regions. *J Proteome Res* 2007; **6**:1899-1916.
7 Xie H, Vucetic S, Iakoucheva LM, *et al*. Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions. *J Proteome Res* 2007; **6**:1882-1898.
8 Xie H, Vucetic S, Iakoucheva LM, *et al*. Functional anthology of intrinsic disorder. 3. Ligands, post-translational modifications, and diseases associated with intrinsically disordered proteins. *J Proteome Res* 2007; **6**:1917-1932.

9 Russell RB, Gibson TJ. A careful disorderliness in the proteome: sites for interaction and targets for future therapies. *FEBS Lett* 2008; **582**:1271-1275.
10 Oldfield CJ, Meng J, Yang JY, *et al*. Intrinsic disorder in protein-protein interaction networks: case studies of complexes involving p53 and 14-3-3. *BMC Genomics* 2008; **9**:S1.
11 Oldfield CJ, Meng J, Yang JY, *et al*. Flexible nets: disorder and induced fit in the associations of p53 and 14-3-3 with their partners. *BMC Genomics* 2008; **9**:S1.
12 Tompa P, Csermely P. The role of structural disorder in the function of RNA and protein chaperones. *Faseb J* 2004; **18**:1169-1175.
13 Spolar RS, Record MT Jr. Coupling of local folding to site-specific binding of proteins to DNA. *Science* 1994; **263**:777-784.
14 Pauling L. A Theory of the structure and process of formation of antibodies. *J Am Chem Soc* 1940; **62**:2643-2657.
15 Fuxreiter M, Simon I, Friedrich P, Tompa P. Preformed structural elements feature in partner recognition by intrinsically unstructured proteins. *J Mol Biol* 2004; **338**:1015-1026.
16 Espinoza-Fonseca LM. Reconciling binding mechanisms of intrinsically disordered proteins. *Biochem Biophys Res Commun* 2009; **382**:479-482.
17 Wright PE, Dyson HJ. Linking folding and binding. *Curr Opin Struct Biol* 2009; **19**:31-38.
18 Dunker AK, Garner E, Guilliot S, *et al*. Protein disorder and the evolution of molecular recognition: theory, predictions and observations. *Pac Symp Biocomput* 1998:473-484.
19 Romero R, Zaidi S, Fang YY, *et al*. Functional profiling by alternative splicing and intrinsic protein disorder. *Proc Natl Acad Sci USA* 2006; **103**:8390-8395.
20 Tompa P. Intrinsically unstructured proteins. *Trends Biochem Sci* 2002; **27**:527-533.
21 Uversky VN. Natively unfolded proteins: a point where biology waits for physics. *Protein Sci* 2002; **11**:739-756.
22 Daughdrill GW, Narayanaswami P, Gilmore SH, Belczyk A, Brown CJ. Dynamic behavior of an intrinsically unstructured linker domain is conserved in the face of negligible amino acid sequence conservation. *J Mol Evol* 2007; **65**:277-288.
23 Dunker AK, Lawson JD, Brown CJ, *et al*. Intrinsically disordered protein. *J Mol Graph Model* 2001; **19**:26-59.
24 Wright PE, Dyson HJ. Intrinsically unstructured proteins: reassessing the protein structure-function paradigm. *J Mol Biol* 1999; **293**:321-331.
25 Weinreb PH, Zhen W, Poon AW, Conway KA, Lansbury PT Jr. NACP, a protein implicated in Alzheimer's disease and learning, is natively unfolded. *Biochemistry* 1996; **35**:13709-13715.
26 Romero P, Obradovic Z, Dunker AK. Natively disordered proteins: functions and predictions. *Appl Bioinformatics* 2004; **3**:105-113.
27 Calvert R, Ungewickell E, Gratzer W. A conformational study of human spectrin. *Eur J Biochem* 1980; **107**:363-367.
28 Veverka V, Henry AJ, Slocombe PM, *et al.* Characterisation of the structural features and interactions of sclerostin: molecular insight into a key regulator of Wnt-mediated bone formation. *J Biol Chem* 2009; **284**:10890-10900.
29 Golovanov AP, Chuang TH, DerMardirossian C, *et al.* Structure-activity relationships in flexible protein domains: regulation of rho GTPases by RhoGDI and D4 GDI. *J Mol Biol* 2001; **305**:121-135.
30 Dastmalchi S, Church WB, Morris MB, Iismaa TP, Mackay JP. Presence of transient helical segments in the galanin-like peptide evident from (1)H NMR, circular dichroism, and prediction studies. *J Struct Biol* 2004; **146**:261-271.
31 Kissinger CR, Parge HE, Knighton DR, *et al*. Crystal structures of human calcineurin and the human FKBP12-FK506-calcineurin complex. *Nature* 1995; **378**:641-644.
32 Radivojac P, Vucetic S, O'Connor TR, *et al*. Calmodulin signaling: analysis and prediction of a disorder-dependent molecular recognition. *Proteins* 2006; **63**:398-410.
33 Sickmeier M, Hamilton JA, LeGall T, *et al*. DisProt: the Database of Disordered Proteins. *Nucleic Acids Res* 2007; **35**:D786-D793.
34 Williams RJ. The conformation properties of proteins in solution. *Biol Rev Camb Philos Soc* 1979; **54**:389-437.
35 Uversky VN, Gillespie JR, Fink AL. Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins* 2000; **41**:415-427.
36 Romero P, Obradovic Z, Kissinger C, Villafranca JE, Dunker AK. Identifying disordered regions in proteins from amino acid sequence. *Proc IEEE Int Conf Neural Networks* 1997; **1**:90-95.
37 Li X, Romero P, Rani M, Dunker AK, Obradovic Z. Predicting protein disorder for N-, C-, and internal regions. *Genome Inform* 1999; **10**:30-40.
38 Romero P, Obradovic Z, Li X, *et al*. Sequence complexity of disordered protein. *Proteins* 2001; **42**:38-48.
39 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.
40 Obradovic Z, Peng K, Vucetic S, Radivojac P, Dunker AK. Exploiting heterogeneous sequence properties improves prediction of protein disorder. *Proteins* 2005; **61**:176-182.
41 Peng K, Radivojac P, Vucetic S, Dunker AK, Obradovic Z. Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics* 2006; **7**:208.
42 Prilusky J, Felder CE, Zeev-Ben-Mordehai T. FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics* 2005; **21**:3435-3438.
43 Linding R, Russell RB, Neduva V, Gibson TJ. GlobPlot: exploring protein sequences for globularity and disorder. *Nucleic Acids Res* 2003; **31**:3701-3708.
44 Linding R, Jensen LJ, Diella F, *et al*. Protein disorder prediction: implications for structural proteomics. *Structure* 2003; **11**:1453-1459.
45 Jones DT, Ward JJ. Prediction of disordered regions in proteins from position specific score matrices. *Proteins* 2003; **53**:573-578.
46 Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol* 2004; **337**:635-645.
47 Ward JJ, McGuffin LJ, Bryson K, Buxton BF, Jones DT. The DISOPRED server for the prediction of protein disorder. *Bioinformatics* 2004; **20**:2138-2139.
48 Bryson K, McGuffin LJ, Marsden RL, *et al*. Protein structure prediction servers at University College London. *Nucleic Ac-*

ids Res 2005; **33**:36-38.
49 MacCallum RM. Order/disorder prediction with self organizing maps. Available from: http://www.forcasp.org/paper2127.html.
50 Dosztányi Z, Csizmók V, Tompa P, Simon I. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J Mol Biol* 2005; **347**:827-839.
51 Dosztanyi Z, Csizmok V, Tompa P, Simon I. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* 2005; **21**:3433-3434.
52 Garbuzynskiy SO, Lobanov MY, Galzitskaya OV. To be folded or to be unfolded. *Protein Sci* 2004; **13**:2871-2877.
53 Galzitskaya OV, Garbuzynskiy SO, Lobanov MY. FoldUnfold: web server for the prediction of disordered regions in protein chain. *Bioinformatics* 2006; **22**:2948-2949.
54 Galzitskaya OV, Garbuzynskiy SO, Lobanov MY. Expected packing density allows prediction of both amyloidogenic and disordered regions in protein chains. *J Phys* 2007; **19**:285225 (15 pp).
55 Yang ZR, Thomson R, McNeil P, Esnouf RM. RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics* 2005; **21**:3369-3376.
56 Cheng J, Sweredoski MJ, Baldi P. Accurate prediction of protein disordered regions by mining protein structure data. *Data Mining and Knowledge Discovery* 2005; **11**:213-222.
57 Su CT, Chen CY, Ou YY. Protein disorder prediction by condensed PSSM considering propensity for order or disorder. *BMC Bioinformatics* 2006; **7**:319.
58 Su CT, Chen CY, Hsu CM. iPDA: integrated protein disorder analyzer. *Nucleic Acids Res* 2007; **35**:465-472.
59 Vullo A, Bortolami O, Pollastri G, Tosatto SC. Spritz: a server for the prediction of intrinsically disordered regions in protein sequences using kernel machines. *Nucleic Acids Res* 2006; **34**:164-168.
60 Ishida T, Kinoshita K. PrDOS: prediction of disordered protein regions from amino acid sequence. *Nucleic Acids Res* 2007; **35**:460-464.
61 Oldfield CJ, Cheng Y, Cortese MS, *et al*. Comparing and combining predictors of mostly disordered proteins. *Biochemistry* 2005; **44**:1989-2000.
62 Bourhis JM, Receveur-Brechot V, Oglesbee M, *et al*. The intrinsically disordered C-terminal domain of the measles virus nucleoprotein interacts with the C-terminal domain of the phosphoprotein via two distinct sites and remains predominantly unfolded. *Protein Sci* 2005; **14**:1975-1992.
63 Bracken C, Iakoucheva LM, Romero PR, Dunker AK. Combining prediction, computation and experiment for the characterization of protein disorder. *Curr Opin Struct Biol* 2004; **14**:570-576.
64 Callaghan AJ, Aurikko JP, Ilag LL, *et al*. Studies of the RNA degradosome-organizing domain of the Escherichia coli ribonuclease RNase E. *J Mol Biol* 2004; **340**:965-979.
65 Longhi S, Receveur-Brechot V, Karlin D, *et al*. The C-terminal domain of the measles virus nucleoprotein is intrinsically disordered and folds upon binding to the C-terminal moiety of the phosphoprotein. *J Biol Chem* 2003; **278**:18638-18648.
66 Bourhis JM, Johansson K, Receveur-Brechot V, *et al*. The C-terminal domain of measles virus nucleoprotein belongs to the class of intrinsically disordered proteins that fold upon binding to their physiological partner. *Virus Res* 2004; **99**:157-167.
67 Kingston RL, Hamel DJ, Gay LS, Dahlquist FW, Matthews BW. Structural basis for the attachment of a paramyxoviral polymerase to its template. *Proc Natl Acad Sci USA* 2004; **101**:8301-8306.
68 Chandran V, Luisi BF. Recognition of enolase in the Escherichia coli RNA degradosome. *J Mol Biol* 2006; **358**:8-15.
69 Liu Y, Matthews KS, Bondos SE. Multiple intrinsically disordered sequences alter DNA binding by the homeodomain of the Drosophila hox protein ultrabithorax. *J Biol Chem* 2008; **283**:20874-20887.
70 Brown CJ, Takayama S, Campen AM, *et al*. Evolutionary rate heterogeneity in proteins with long disordered regions. *J Mol Evol* 2002; **55**:104-110.
71 Uversky VN, Dunker AK. Intrinsically disordered proteins: controlled chaos. *Science* 2008; **322**: 1340-1341
72 Gsponer J, Futschik ME, Teichmann SA, Babu MM. Tight regulation of unstructured proteins: from transcript synthesis to protein degradation. *Science* 2008; **322**:1365-1368.
73 Bandaru V, Cooper W, Wallace SS, Doublie S. Overproduction, crystallization and preliminary crystallographic analysis of a novel human DNA-repair enzyme that recognizes oxidative DNA damage. *Acta Crystallogr D Biol Crystallogr* 2004; **60**:1142-1144.
74 Galea CA, High AA, Obenauer JC, *et al*. Large-scale analysis of thermostable, mammalian proteins provides insights into the intrinsically disordered proteome. *J Proteome Res* 2009; **8**:211-226.
75 Cortese MS, Baird JP, Uversky VN, Dunker AK. Uncovering the unfoldome: enriching cell extracts for unstructured proteins by acid treatment. *J Proteome Res* 2005; **4**:1610-1618.
76 Galea CA, Pagala VR, Obenauer JC, *et al.* Proteomic studies of the intrinsically unstructured mammalian proteome. *J Proteome Res* 2006; **5**:2839-2848.
77 Szollosi E, Bokor M, Bodor A, *et al.* Intrinsic structural disorder of DF31, a Drosophila protein of chromatin decondensation and remodeling activities. *J Proteome Res* 2008; **7**:2291-2299.
78 Tress ML, Bodenmiller B, Aebersold R, Valencia A. Proteomics studies confirm the presence of alternative protein isoforms on a large scale. *Genome Biol* 2008; **9**:R162.
79 Bourhis JM, Canard B, Longhi S. Predicting protein disorder and induced folding: from theoretical principles to practical applications. *Curr Protein Pept Sci* 2007; **8**:135-149.
80 Dosztányi Z, Sándor M, Tompa P, Simon I. Prediction of protein disorder at the domain level. *Curr Protein Pept Sci* 2007; **8**:161-171.
81 Ferron F, Longhi S, Canard B, Karlin D. A practical overview of protein disorder prediction methods. *Proteins* 2006; **65**:1-14.
82 Kryshtafovych A, Fidelis K, Moult J. Progress from CASP6 to CASP7. *Proteins* 2007; **69**:194-207.
83 Melamud E, Moult J. Evaluation of disorder predictions in CASP5. *Proteins* 2003; **53**:561-565.
84 Jin Y, Dunbrack RL Jr. Assessment of disorder predictions in

CASP6. *Proteins* 2005; **61**:167-175.
85 Bordoli L, Kiefer F, Schwede T. Assessment of disorder predictions in CASP7. *Proteins* 2007; **69**:129-136.
86 Gunnasekaran K, Tsai CJ, Nussinov R. Analysis of ordered and disordered protein complexes reveals structural features discriminating between stable and unstable monomers. *J Mol Biol* 2004; **341**:1327-1341.
87 Cheng Y, Oldfield CJ, Meng J, *et al*. Mining alpha-helix-forming molecular recognition features with cross species sequence alignments. *Biochemistry* 2007; **46**:13468-13477.
88 Oldfield CJ, Cheng Y, Cortese MS, *et al*. Coupled folding and binding with alpha-helix-forming molecular recognition elements. *Biochemistry* 2005; **44**:12454-12470.
89 Romero P, Obradovic Z, Dunker AK. Sequence data analysis for long disordered regions prediction in the calcineurin family. *Genome Inform* 1997; **8**:110-124.
90 Dunker AK, Obradovic Z, Romero P, Garner EC, Brown CJ. Intrinsic protein disorder in complete genomes. *Genome Inform* 2000; **11**:161-171.
91 Weathers EA, Paulaitis ME, Woolf TB, Hoh JH. Reduced amino acid alphabet is sufficient to accurately recognize intrinsically disordered protein. *FEBS Lett* 2004; **576**:348-352.
92 Thomas PD, Dill KA. An iterative method for extracting energy-like quantities from protein structures. *Proc Natl Acad Sci USA* 1996; **93**:11628-11633.
93 Galaktionov SG, Marshall GR. Technical report: Prediction of Protein Structure in Terms of Intraglobular Contacts: 1D to 2D to 3D. Institute for Biomedical Computing, Washington University, St. Louis, 1996.
94 Hecker J, Yang JY, Cheng J. Protein disorder prediction at multiple levels of sensitivity and specificity. *BMC Genomics* 2008; **9**:S9.
95 Ferron F, Longhi S, Henrissat B, Canard B. Viral RNA-polymerases – a predicted 2'-O-ribose methyltransferase domain shared by all Mononegavirales. *Trends Biochem Sci* 2002; **27**:222-224.
96 Karlin D, Ferron F, Canard B, Longhi S. Structural disorder and modular organization in Paramyxovirinae N and P. *J Gen Virol* 2003; **84**:3239-3252.
97 Ferron F, Rancurel C, Longhi S, *et al*. VaZyMolO: a tool to define and classify modularity in viral proteins. *J Gen Virol* 2005; **86**:743-749.
98 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**:3389-3402.
99 Corpet F. Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res* 1988; **16**:10881-10890.
100 Callebaut I, Labesse G, Durand P, *et al*. Deciphering protein sequence information through hydrophobic cluster analysis (HCA): current status and perspectives. *Cell Mol Life Sci* 1997; **53**:621-645.
101 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**:2337-2338.
102 Hirose S, Shimizu K, Kanai S, Kuroda Y, Noguchi T. POODLE-L: a two-level SVM prediction system for reliably predicting long disordered regions. *Bioinformatics* 2007; **23**:2046-2053.
103 Yang MQ, Yang JY. IUP: intrinsically unstructured protein predictor – a software tool for analyzing polypeptide sequences. BioInformatics and BioEngineering, 2006. BIBE 2006. Sixth IEEE Symposium on 16-18 Oct.:3-11.
104 Schlessinger A, Punta M, Rost B. Natively unstructured regions in proteins identified from contact predictions. *Bioinformatics* 2007; **23**:2376-2384.
105 Shimizu K, Muraoka Y, Hirose S, Tomii K, Noguchi T. Predicting mostly disordered proteins by using structure-unknown protein data. *BMC Bioinformatics* 2007; **8**:78.
106 Bulashevska A, Eils R. Using bayesian multinomial classifier to predict where a given protein sequence is intrinsically disordered. *J Theor Biol* 2008; **254**:799-803.
107 Wang L, Sauer UH. OnD-CRF: predicting order and disorder in proteins using conditional random fields. *Bioinformatics* 2008; **24**:1401-1402.
108 Xue B, Oldfield CJ, Dunker AK, Uversky VN. CDF it all: consensus prediction of intrinsically disordered proteins based on various cumulative distribution functions. *FEBS Lett* 2009; **583**:1469-1474.
109 Ishida T, Kinoshita K. Prediction of disordered regions in proteins based on the meta approach. *Bioinformatics* 2008; **24**:1344-1348.
110 Schlessinger A, Punta M, Yachdav G, Kajan L, Rost B. Improved disorder prediction by combination of orthogonal approaches. *PLoS ONE* 2009; **4**:e4433.
111 McGuffin LJ. Intrinsic disorder prediction from the analysis of multiple protein fold recognition models. *Bioinformatics* 2008; **24**:1798-1804.
112 Lieutaud P, Canard B, Longhi S. MeDor: a metaserver for predicting protein disorder. *BMC Genomics* 2008; **9**:S25.
113 Chandonia JM, Karplus M. New methods for accurate prediction of protein secondary structure. *Proteins* 1999; **35**:293-306.
114 Chandonia JM. StrBioLib: a Java library for development of custom computational structural biology applications. *Bioinformatics* 2007; **23**:2018-2020.
115 Coeytaux K, Poupon A. Prediction of unfolded segments in a protein sequence based on amino acid composition. *Bioinformatics* 2005; **21**:1891-1900.
116 Obradovic Z, Peng K, Vucetic S, *et al*. Predicting intrinsic disorder from amino acid sequence. *Proteins* 2003; **53** Suppl 6:566-572.
117 Kall L, Krogh A, Sonnhammer EL. A combined transmembrane topology and signal peptide prediction method. *J Mol Biol* 2004; **338**:1027-1036.
118 Daughdrill GW, Pielak GJ, Uversky VN, Cortese MS, Dunker AK. Natively disordered proteins. In: Buchner J, Kiefhaber T, eds. *Handbook of Protein Folding*. Weinheim, Germany: Wiley-VCH, Verlag GmbH & Co., 2005:271-353.
119 Vucetic S, Brown CJ, Dunker AK, Obradovic Z. Flavors of protein disorder. *Proteins* 2003; **52**:573-584.
120 Uversky VN, Oldfield CJ, Dunker AK. Intrinsically disordered proteins in human diseases: introducing the D2 concept. *Annu Rev Biophys* 2008; **37**:215-246.
121 Receveur-Brechot V, Bourhis JM, Uversky VN, Canard B, Longhi S. Assessing protein disorder and induced folding. *Proteins* 2006; **62**:24-45.
122 Liu J, Rost B. NORSp: predictions of long regions without

regular secondary structure. *Nucleic Acids Res* 2003; **31**:3833-3835.
123 Gu J, Gribskov M, Bourne PE. Wiggle-predicting functionally flexible regions from primary sequence. *PLoS Comput Biol* 2006; **2**:e90.
124 Vacic V, Uversky VN, Dunker AK, Lonardi S. Composition profiler: a tool for discovery and visualization of amino acid composition differences. *BMC Bioinformatics* 2007; **8**:211.
125 Schlessinger A, Liu J, Rost B. Natively unstructured loops differ from other loops. *PLoS Comput Biol* 2007; **3**:e140.
126 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**:956-963.
127 Sethi D, Garg A, Raghava GP. DPROT: prediction of disordered proteins using evolutionary information. *Amino Acids* 2008; **35**:599-605.
128 Yang JY, Yang MQ. Identification of intrinsically unstructured proteins using hierarchical classifier. *Int J Data Min Bioinform* 2008; **2**:121-133.
129 Han P, Zhang X, Feng ZP. Predicting disordered regions in proteins using the profiles of amino acid indices. *BMC Bioinformatics* 2009; **10**:S42.
130 Han P, Zhang X, Norton RS, Feng ZP. Large-scale prediction of long disordered regions in proteins using random forests. *BMC Bioinformatics* 2009; **10**:8.