BRIEFINGS IN BIOINFORMATICS. VOL 11. NO 2. 225–243
Advance Access published on 10 December 2009
doi:10.1093/bib/bbp061

Bioinformatical approaches to characterize intrinsically disordered/unstructured proteins

Zsuzsanna Dosztányi, Bálint Mészáros and István Simon

Submitted: 9th October 2009; Received (in revised form): 11th November 2009

Abstract

Intrinsically disordered/unstructured proteins exist without a stable three-dimensional (3D) structure as highly flexible conformational ensembles. The available genome sequences revealed that these proteins are surprisingly common and their frequency reaches high proportions in eukaryotes. Due to their vital role in various biological processes including signaling and regulation and their involvement in various diseases, disordered proteins and protein segments are the focus of many biochemical, molecular biological, pathological and pharmaceutical studies. These proteins are difficult to study experimentally because of the lack of unique structure in the isolated form. Their amino acid sequence, however, is available, and can be used for their identification and characterization by bioinformatic tools, analogously to globular proteins. In this review, we first present a small survey of current methods to identify disordered proteins or protein segments, focusing on those that are publicly available as web servers. In more detail we also discuss approaches that predict disordered regions and specific regions involved in protein binding by modeling the physical background of protein disorder. In our review we argue that the heterogeneity of disordered segments needs to be taken into account for a better understanding of protein disorder.

Keywords: *protein disorder; coupled folding and binding; machine-learning algorithm; prediction method; binary classification*

INTRODUCTION

Classical biochemical studies reinforced the view that the formation of a well-formed structure is a prerequisite for a protein to carry out its function. Following the advice of Crick: 'If you want to understand the function, study the structure,' this notion motivated a large number of structure–function studies and lead to the structure determination of more than 50 000 proteins [1]. While databases of known protein structures have grown relatively slowly, the number of sequences and data on protein interactions increased drastically as a result of new experimental techniques and large-scale sequencing projects. This new information reshaped our view of the protein world [2]. It has become evident that a large number of naturally occurring proteins do not require a well-folded structure to fulfill their biological role [3–6]. These intrinsically unstructured/disordered proteins (IUPs/IDPs) exist as ensembles of rapidly interconverting conformations, even under physiological conditions. Their importance is also underlined by their expected high frequency in proteomes. Using bioinformatic predictors it was estimated that 30–50% of eukaryotic proteins contain at least one long disordered segment [7, 8]. These proteins participate in important regulatory functions in

Corresponding Author. István Simon, Institute of Enzymology, Biological Research Center, Hungarian Academy of Sciences, P.O. Box 7, H-1518 Budapest, Hungary. Tel:+36-1-4669276; Fax: +36-1-4665465; E-mail: simon@enzim.hu

Zsuzsanna Dosztányi is a senior scientist at the Institute of Enzymology. Her current interest focuses on understanding and predicting various structural and functional properties of intrinsically disordered proteins.

Bálint Mészáros is a PhD student at the Institute of Enzymology, his main research interests are disordered proteins and their role in protein–protein interactions.

István Simon is the head of the Protein Research Group at the Institute of Enzymology. His main interest is studying the structure and stability of globular, transmembrane and unstructured proteins by theoretical and computational approaches.

the cell including transcription, translation [9, 10] and cell signaling [2, 11–14]. Several IDPs were shown to be associated with various diseases such as cancer and neurodegenerative diseases [15, 16]. Recently, new strategies for drug discovery were suggested specifically targeting disordered proteins [17]. Recognizing the relevance of these proteins stimulated more systematic efforts aiming at their structural characterization and determination of their mechanisms of action.

From structural point of view, disordered segments are heterogeneous and affect various levels of protein structure. Some of them exist in the form of random-coils that corresponds to a largely random distribution of conformations dominated by extended structures [18]. In reality, however, no protein is ever random-coil, and the macroscopic properties compatible with random-coils do not exclude the possibility of transient short- or long-range interactions resulting in transient structural elements [12, 19]. Disordered proteins can also exist as molten-globules exhibiting a compact but disordered state with some secondary-structure content [4, 18]. In general, while some proteins appear fully disordered, many proteins are composed of both ordered and disordered regions of various lengths [20]. Larger proteins are usually segregated into multiple domains and contain flexible linker regions connecting the domains. Disordered segments can correspond to loop regions or flexible termini within the context of globular proteins [21]. Generally, various types of disorder and the transition between these states can be linked to specific function of the proteins.

In the case of many proteins, like entropic chains, the function directly originates from their permanent disordered state [22]. However, numerous examples show that disordered regions can undergo a disorder-to-order transition upon binding to other macromolecules [23]. Many such disordered proteins are involved in molecular recognition and function via binding to a structured protein partner. The inherent flexibility of disordered proteins imposes specific thermodynamic and kinetic properties during complex formation and allows them to combine low affinity with high specificity in their binding [2, 24]. In the bound form, they can lend themselves to traditional structure determination. Although the number of such complexes is rather limited, these examples show significant differences compared to the complexes formed between globular proteins [25–27]. This indicates that the distinct properties of disordered protein regions are imprinted even in the rigid conformation adopted in the complex. As exemplified by the C-terminal region of p53, they can bind multiple partners even in different conformations [28], and conformational preferences observed in their free form is not necessarily indicative of the adopted conformation in the bound state [23]. Disordered proteins often use a single continuous segment for partner binding whereas the binding sites of ordered proteins are more segmented [26, 27]. Beside their involvement in protein–protein interactions, these proteins are also subjects of various post-translational modifications that control their functions, localization and turnover [12, 29].

The detailed structural and functional characterization of disordered proteins is quite a challenging task [30, 31]. The existing experimental procedures are highly biased for ordered proteins, and most techniques provide only indirect information about disorder [2]. Consequently, the current list of experimentally verified disordered proteins is rather limited. Because of these difficulties, bioinformatic tools play a very important role in the identification and characterization of IDPs.

In the past few decades several algorithms were developed to predict various aspects of proteins with unique structures, including prediction of secondary-structure [32–34], solvent accessibility of residues [35, 36], the covalent state of residue Cys [37–39], domain boundaries of globular proteins [40–42], sites for protein interactions [43], as well as the topology of transmembrane proteins [44–46]. These methods generally use only sequence information as input that is also available for unstructured proteins. The success of these methods suggests that appropriate tools can be developed analogously to study proteins that have no unique structure. In fact, several methods have been developed to identify unstructured proteins or protein segments and a few additional methods are also available to estimate further properties of these unstructured segments [47–49]. In this work we highlight some of the main ideas and challenges in predicting protein disorder in general. We provide a brief overview of current methods focusing on those that are publicly available. We put a special emphasis on methods that can provide more specific information about disordered proteins.

THE BASIC COMPONENTS OF DISORDER PREDICTION METHODS

Databases

An essential components of disorder prediction methods are the various datasets used for optimization and evaluation. Ordered proteins are collected from the Protein Data Bank (PDB) [1, 50], as the presence of known coordinates is generally accepted as a direct evidence of structural order. Indirectly, the PDB also contains information about disordered regions. In protein structures solved by X-ray crystallography, disorder is defined by missing electron density. Missing residues usually appear within the context of ordered structures, either as terminal regions or short loops within an otherwise ordered protein. Their length spans from a single residue to hundreds, but most often these regions are <30 residues long. The current most comprehensive resource on protein disorder is the DisProt database [51]. It aims to collect disordered proteins and protein regions characterized by various experimental techniques. Most of these regions are more than 30 residues long. However, our current collection of experimentally well-characterized disordered protein regions is still rather limited. To overcome this limitation, large sequence databases can also be exploited in certain type of predictions [52]. These databases are likely to contain many more disordered proteins that are yet uncharacterized, but they are not biased by limitations of current experimental techniques.

Datasets of experimentally verified ordered and disordered regions contain many mis-classified segments. Some regions that appear ordered in the crystal structure only adopt a well-defined structure as part of a larger complex, but would be disordered in isolation. Disordered segments are even more prone to misclassification since most longer disordered regions are characterized by semi-quantitative experiments that lack position specific information. Furthermore, the order/disorder status can also be sensitive to various environmental conditions [53, 54]. The lack of sufficiently large datasets and the noise in the assignment of order and disorder represent a serious limitation in developing accurate prediction methods for protein disorder.

Evaluation

The performance of various disorder prediction methods was critically assessed in the last four rounds of CASP experiments [55–58]. The reports of these meetings also provide a guide for the evaluation of various disorder prediction methods [56]. According to one of the favored evaluation measures, the area under the ROC curve (AUC), top methods can reach at least 0.9 AUC. In other terms, they can identify ~70% of disordered residues at the expense of misclassifying <10% of ordered residues. However, CASP evaluations are restricted to residues with missing X-ray coordinates and there is no similar blind testing for long disordered regions. Several authors carried out performance tests of various methods specifically on longer segments of ordered and disordered proteins and found ~80% efficiency on both datasets [20, 59–61]. However, these results can be biased and should be treated with a grain of salt. The general wisdom is that the performance of disorder prediction methods critically depends on the dataset used for testing, or more generally, the type of disorder studied. It is also influenced by the evaluation criteria. For this reason, we do not try to rank various prediction methods. Rather, we focus on the key concepts and ideas in the field of disorder prediction.

Basic sequence properties of disordered segments

The first analyses of sequences of disordered proteins revealed significant differences in the amino acid composition of ordered and disordered proteins. Disordered proteins are generally depleted in bulky hydrophobic and aromatic amino acids and are enriched in polar and charged amino acids. At closer inspection, however, various datasets of disordered protein sequences exhibited further variations in their sequential bias. Differences could be observed depending on the experimental method used to identify disordered regions (e.g. CD, NMR, or X-ray crystallography) [62], depending on the length of disordered regions [63], and the location in the sequence (N- and C-terminal, middle regions) [64]. Although these differences are smaller compared to the differences observed between ordered and disordered proteins, they should be taken into account during the development of prediction methods.

The amino acid compositional bias of disordered proteins suggests the relevance of hydrophobicity scales for the discrimination of ordered and disordered segments. Among various amino acid scales, properties related to flexibility and coordination number had the highest discriminatory power [65, 66]. Several disordered prediction methods are based on simple amino acid propensity scales [6, 18, 67]. Globplot is based on the hypothesis that the tendency for disorder can be expressed as the difference of the amino acid propensities to be in coil and regular secondary-structure elements [68]. A specific amino-acid scale optimized to discriminate ordered and disordered regions was also constructed [66].

The appeal of single amino acid propensities is that they are easy to calculate and to interpret, however, they are limited to a single effect. This can be insufficient to account for the complex phenomenon of protein disorder. Such properties, however, are also useful to reduce the dimensionality of the input data. By focusing on the relevant properties, an increased performance can be achieved during prediction. Several methods exploited amino-acid scales in their predictions, including PONDR VL-XT [64], and VSL2 [59] or DisPSSMP [69].

Low complexity

The most traditional approach to filter out non-globular protein regions relies on finding low-complexity segments [70, 71]. Compositional complexity is a bioinformatic measure of the randomness of a protein sequence and it is calculated based on sequence entropy. It was first introduced for the purpose of sequence alignments and searches, to filter out regions which violated the basic assumption of the underlying statistical model of sequence alignments. Low complexity segments are common in disordered proteins, and the more biased the amino acid composition of disordered segments, it is more likely to be also of low complexity [72]. Nevertheless, many disordered proteins are practically indistinguishable from ordered proteins based on their sequence complexity alone, while low complexity regions can also include ordered structural proteins or proteins with strong structural propensity, like collagens, coiled-coils or other fibrous proteins. Therefore, low complexity in itself is not sufficient to identify protein disorder in general, although, it can capture an important component of certain types of protein disorder [72, 73].

Evolutionary information

Using amino acid profiles calculated from evolutionary related sequences [74] instead of a single sequence window as an input has significantly boosted various areas of structure prediction. The incorporation of this information into prediction of protein disorder, however, is more problematic. Disordered regions are often excluded prior to sequence searches by using low complexity filter. Some remaining regions possess a strong sequence bias that can still distort the result of sequence similarity searches. Compared to globular proteins, both the type and the rate of amino acid substitutions differ in the case of disordered proteins [75]. Generally, most disordered proteins evolve faster due to the lack of structural constraints. In certain cases, amino acid sequence conservation is not required for the conservation of dynamic behavior and presumably molecular function of disordered regions [76]. Nevertheless, evolutionary conserved IDPs can also be found, especially among those that are involved in complex formation [26, 75]. Generally, the incorporation of evolutionary information led to a much smaller increase in the performance of disorder prediction methods, compared for example to secondary structure prediction methods [59].

Secondary structure and disorder

Bioinformatical methods often benefited from additional predicted properties, including secondary structure or solvent accessibility [32, 35]. These predicted properties can also be exploited in the prediction of protein disorder. It should be noted, however, that these methods have been exclusively trained on ordered proteins, and should be used only with caution outside this realm. For example, predicted secondary-structure does not necessarily contradict protein disorder. Often these regions correspond to transient secondary-structural elements, or to the conformation adopted in the complex form [77]. In the isolated form, with the exception of highly specific scenarios [78], predicted secondary-structures are not expected to be stable for disordered proteins. Nevertheless, several methods include prediction of secondary-structure or solvent accessibility in their input [79–81]. Furthermore, certain types of disordered proteins can be identified as long regions with no predicted secondary-structures [82].

OVERVIEW OF PREDICTION METHODS FOR DISORDERED REGIONS

Machine learning approaches

The prediction of protein disorder can be viewed as a classic binary classification problem and can be addressed by standard machine learning techniques. The underlying assumption is that sequence features calculated from a local sequence window can be directly mapped into the property of order or disorder. Most methods assign disordered and ordered status at the amino acid residue level. The novelty of many disorder prediction methods based on machine learning approaches lies in the representation of input information, rather than in the algorithms themselves. A comprehensive review [49] of published methods appeared in the literature recently. Here, we focus on those methods which are publicly available via web servers or standalone programs, and provide residue based predictions. A summary of these methods can be found in Table 1.

The first method developed for the prediction of disordered proteins is PONDR VL-XT [72]. This method is based on feed-forward neural networks, one the most common methods in the field of bioinformatics. PONDR VL-XT is composed of two separate predictors developed for the N- and C-terminal regions trained on terminal disordered regions characterized by X-ray [64] and a specific predictor for middle regions trained on variously characterized long disordered regions [83]. The inputs of these predictors are specific sequence attributes calculated within a given window. These attributes include the coordination number, net charge, hydropathy, and the fraction of various amino acid groups. Because the available datasets at the time were very small, a small number of attributes were selected by analyzing their discriminatory power, their orthogonality, and based on their effect on the performance. The resulting method was found particularly useful to pinpoint certain regions that are candidates for undergoing disorder-to-order transitions [84, 85].

Several other methods use standard feed-forward neural networks. One of these is PONDR VL3 [86] that was trained on a much larger dataset of variously characterized disordered segments, compared to VL-XT. The input is formed by 18 amino acid frequencies, the average flexibility and sequence complexity, calculated within a window of 41 residues. DisEMBL [87] was trained specifically on missing residues of X-ray structures but it also incorporates additional methods to predict residues with high B-factor and in loop regions. The differences in these three predictors, trained on missing residues, high B-factor regions and loops, respectively, underlined the distinct features of each group.

A more recent method, DisPSSMP [69] used radial basis function networks as a training algorithm. Although it uses position specific scoring matrices generated by PSI-BLAST [74], these matrices are significantly condensed using basic physico-chemical properties. The optimal set of properties is the result of a step-wise feature selection procedure. The newer version of the method incorporates secondary-structure predictions as well [80]. Using a different approach, RONN predictions [88] are based on similarity to known examples of disordered segments. In this method, sub-sequences of a query sequence are aligned to all prototype segments, and the similarity to these sequence fragments is calculated using a standard mutation matrix. The resulting homology scores are combined by a modified version of radial basis function network called bio-basis function neural networks.

Another class of standard machine-learning algorithms is support vector machines (SVMs). SVMs are less prone to overfitting, compared to neural networks, and can be trained more efficiently. The first method utilizing SVMs for the prediction of disorder was DISOPRED2 [8]. This method was trained on a dataset of missing residues of solved structures, separately for N-, C- and middle regions. The input was generated from PSI-BLAST generated profiles of position specific scoring matrices [74]. One of the advantages of SVMs is that it can incorporate greater cost of misclassification for one of the classes, therefore it can compensate for unbalanced datasets. This is the key for the low false positive rate of DISOPRED2. PrDOS [89] is also a basic SVM based prediction method, however, it is combined with a template based prediction that takes into account the disordered status observed in structures homologous to the query sequence.

The POODLE-S [90] and POODLE-L [61] predictors are specific methods for recognizing short and long disordered regions. They employed SVMs with radial basis kernels for training and constructed the input from phsyico-chemical properties using

Table 1: Summary of the 13 analyzed disorder-prediction methods

Name of predictor	URL	Training dataset	Algorithm	Input data of the algorithm
VL-XT [72]	Not publicly available	XT: missing residues in X-ray structures (terminal regions)VL: variously characterized long disordered segments	Neural network	Amino acid frequencies, amino acid propensities
VL3 [86]	http://www.ist.temple.edu/disprot/Predictors.html	Variously characterized long disordered segments	Neural network	Amino acid frequencies, amino acid propensities, sequence complexity
DisEMBL [87]	http://dis.embl.de/	Missing residues in X-ray structures	Neural network	Single sequence window
DisPSSMP [69]	http://biominer.bime.ntu.edu.tw/ipda/	DisProt	Radial basis function neural network	PSI-BLAST PSSM condensed by physico-chemical properties, secondary-structure prediction
RONN [88]	http://www.strubi.ox.ac.uk/RONN/	Missing residues in X-ray structures	Bio-basis function neural network	Single sequence window
DISOPRED2 [8]	http://bioinf.cs.ucl.ac.uk/disopred/disopred.html	Missing residues in X-ray structures	SVM and neural network	PSI-BLAST PSSM windows
PrDOS [89]	http://prdos.hgc.jp/cgi-bin/top.cgi	Missing residues in X-ray structures	SVM and template-based prediction	PSI-BLAST and homologous structures
DISpro [79]	http://scratch.proteomics.ics.uci.edu/	Missing residues in X-ray structures	ID recursiveneural network	Full length PSI-BLAST PSSM,secondary structure and solvent accessibility prediction
OnD-CRF [81]	http://babel.ucmp.umu.se/ond-crf/	Missing residues in X-ray structures	Conditional random fields	Single sequence, secondary-structure prediction
DRIP-PRED [91]	http://www.sbc.su.se/~maccallr/disorder/	UniProt sequences	Kohonen SOM	PSI-BLAST PSSM windows, secondary-structure prediction
VSL2B [59,93]	http://www.ist.temple.edu/disprot/Predictors.html	Missing residues in X-ray structures and DisProt	SVM	Amino acid propensities, sequence complexity, (PSI-BLAST PSSM) (secondary-structure prediction)
POODLE-I [94]	http://mbs.cbrc.jp/poodle/poodle.html	Missing residues in X-ray structures, DisProt and SwissProt	SVM and SGT	Amino acid propensities, PSI-BLAST PSSM, secondary-structure prediction
IUPred [100]	http://iupred.enzim.hu/	None	Biophysical model	Amino acid composition

SVM, support vector machine; SOM, self-organizing map; SGT, spectral graph transducer; PSSM, position specific scoring matrix. Column 3 shows the dataset for disorder on which the methods were trained, column 4 shows the basic implemented algorithm and column 5 shows the quantities the algorithm uses to calculate the final prediction score.

PSI-BLAST profiles. In the case of POODLE-L, which aims to predict disordered segments longer than 40 residues, the input is composed of mean hydropathy, average contact density propensity, mean net charge, sequence complexity, amino acid compositions relative to the composition of disordered and ordered training sets and secondary structure preferences. POODLE-S calculates the input vector by using physico-chemical features and a reduced amino acid set of position-specific scoring matrices.

In the case of feed forward neural networks and SVMs the prediction for each residue is independent of the prediction of other residues. In contrast, recurrent networks can also propagate data from later processing stages to earlier stages. Such technique is used in DISpro [79]. It employs a 1D recursive neural network and leverages evolutionary information as well as predicted secondary structure and solvent accessibility. Instead of using a fixed window size, the prediction at each position depends on the entire sequence through a recursive network of neighboring positions. The recently published method OnD-CRF [81] utilizes conditional random fields for the prediction of protein disorder. Common to both methods is the ability to take into account the predicted disorder tendency of neighboring positions.

Instead of using explicit datasets for disordered proteins, methods can also exploit the information stored in large sequence databases. The DRIP-PRED method [91] uses this strategy. For this purpose, sequence profile windows corresponding to the complete database of UniProt sequences were clustered using Kohonen's self-organizing map [91]. It was found that there are regions of 'UniProt space' which are essentially unpopulated by proteins of known structure. Sequence windows which map to these locations are not well represented in the PDB and therefore are predicted as disordered. In a different approach, mostly ordered and disordered proteins are classified by spectral graph transducer by POODLE-W [52]. This method uses sequences of the SwissProt database which is more reliable compared to the UniProt database. These sequences are treated as an unlabeled dataset and are used together with ordered proteins collected from the PDB during training. A graph is constructed based on the similarity of both labeled and unlabeled sequences, calculated from their amino acid composition. The classification is based on the optimal separation of this graph into ordered and disordered data.

Generally, the prediction methods assign a score to each residue in the sequence. These scores, however, can show large fluctuations from one residue to the other. This effect is generally smoothed out by various techniques. The simplest solution uses the average of the scores calculated within a given window. The optimal window size is characteristic of the disordered data, for example mostly disordered proteins, or longer segments of disorder prefer larger windows for smoothing. An alternative approach for smoothing is based on curve fitting algorithms, applied for example in the case of GlobPlot [68]. As the aim of GlobPlot approach is to identify ordered domains, it also eliminates short segments predicted as ordered that are too short to fold on their own, and a similar filter is applied for disordered segments. Several methods apply a second level of prediction using the output of the first level prediction as an input. DISOPRED2 employs a neural network based predictor, while POODLE-L implemented another SVM based predictor for this purpose. There are also specialized predictors to assign disorder and order status at the level of whole proteins [92].

The methods described so far are all specific to one type of protein disorder only, represented either by DisProt [51] dataset or missing residues of X-ray structures. Their performance tested on the other dataset resulted in significantly lower efficiencies. This problem was first addressed by the PONDR VSL2 method [59, 93]. It is composed of two separate predictors optimized for short and long (>30 residues) disordered regions that are combined by an independent meta-predictor. Linear SVM was chosen as the learning algorithm, because it has similar performance but better generalization ability compared to other techniques. The input of all three methods are composed of various amino acid propensities, sequence complexity, and optionally sequence profiles and secondary-structure predictions, calculated within a sliding local window. At the first level, the two methods predict short and long disordered segments, respectively. The meta-predictor then determines the optimal weight to combine the output of these two composite predictors. This architecture ensured that PONDR VSL2 has a more balanced performance on disordered segments of various lengths. Another approach, POODLE-I [94] also integrates methods that target different disordered regions according to their length, by incorporating their specific predictors that recognize short and long disordered segments as well as mostly disordered proteins, and combine their outputs with various secondary structure predictions.

Meta approaches that integrate the results of several prediction methods have been very successful in various areas of structure predictions [95] and appeared for the prediction of protein disorder as well. These methods achieve improved performance by decreasing the noise of individual predictors. Furthermore, since individual disorder prediction methods are often specific to certain types of protein disorder, their combination could cover more aspects of disorder. In this vein, MD combines four, but largely orthogonal predictions [60]. This meta-predictor was trained on the DisProt dataset using a neural network. Othogonality is also a key in the DISOclust server [96]. It improves on the predictions provided by DISOPRED2 by complementing it with structural variability calculated over multiple fold recognition models. The premise of this method is that residues that are highly variable in the 3D space from one model to the other may coincide with regions of disorder. Another method, metaPrDOS [97] predicts disordered regions by integrating the results of eight different prediction methods that are combined into a single predictor using an SVM trained on missing residues. In the GSMetaDisorder server 13 individual methods are combined with prediction of secondary structure and solvent accessibility using a neural network that was trained on data both from DisProt database and missing residues of X-ray structures [98].

The last round of CASP experiment was clearly dominated by meta-predictors [58]. Nevertheless, there is still an urgent need for specialized predictors that can accurately capture certain types of disorder. Although these predictors might be inferior to meta-predictors in certain evaluations, they provide more insight into the structural and even the functional properties of disordered regions.

Physics-based methods

An alternative approach to various machine learning algorithms in predicting protein disorder is the direct implementation of physical principles governing the process of protein folding. It was suggested that disordered proteins can be identified based on the combination of low hydrophobicity and high net charge [6, 18]. The rationale behind this approach is that high net charge leads to charge–charge repulsion and low hydrophobicity means less driving force for a compact structure. This algorithm was implemented in the FoldIndex algorithm to provide position specific prediction [99]. A similar concept is behind the FoldUnfold method [67]. It predicts protein disorder based on the expected average number of contacts per residue. These values are taken from a single amino acid propensity scale that encodes the average number of contacts for the 20 amino acid residues in a dataset of globular proteins. The IUPred algorithm captures the essential cause of protein non-folding in a more general way: if a residue in a protein is not able to form enough favorable intrachain contacts, it will not adopt a stable position in the 3D structure of the chain [100, 101]. If such residues are clustered along a segment of a protein or the whole protein, then this segment or the entire protein will be disordered.

The implementation of the above principle relies on statistical potentials [102]. Statistical potentials are calculated from the observed frequency of interactions between amino acids based on the Boltzmann hypothesis. With these energy-like quantities, the total pairwise energy of a protein can be calculated in a given conformation. The core of IUPred is a method that enables the direct estimation of the interaction energies using the protein sequence alone. In this approach, the estimated energy for each residue depends on the amino acid type but also on the amino acid composition of the sequential neighborhood. Generally, residues with less favorable predicted energies are more likely to be disordered.

The parameters of this method are derived from a globular protein dataset without the use of specific datasets of disordered proteins. As globular protein datasets are considerably larger than that of disordered proteins, this grants the method substantial stability compared to methods where a large number of parameters are trained on a limited and sometimes ambiguous disordered protein dataset.

Prediction of disordered binding regions

Many disordered proteins carry out important functions via binding to other macromolecules that involves coupled folding and binding. Due to their specific functional and structural properties, these binding regions have distinct properties compared to both globular and disordered proteins in general.

While there are many algorithms for predicting IDPs, apparently the choice of methods for predicting regions undergoing disorder-to-order transition upon protein binding is rather limited.

A recent method for the prediction of disordered binding regions, ANCHOR aims to capture the basic biophysical properties of disordered binding segments [103]. The essential feature of these regions is that they exist in a disordered state in isolation, but they can favorably interact with a globular protein and adopt a rigid conformation upon binding. Based on this model, the combination of the high disordered tendency of the environment, the unfavorable intrachain interaction energies and high energetic gain by interacting with a globular protein partner indicates the presence of a disordered binding region. The implementation of these principles follows the basic idea behind IUPred, and are quantified with the use of estimated energies.

The ANCHOR predictor could recognize 68% of disordered binding regions at a segment level, while falsely predicting only 5% of residues in ordered proteins. As the available dataset for experimentally verified disordered protein complexes is limited in size, the benefit of using physical models instead of machine learning algorithms is evident. Another strength of ANCHOR comes from the fact, that the efficiency of the prediction is largely independent of the amino acid composition of the query protein. For example, basic binding regions, such as certain calmodulin binding sites, are recovered with approximately the same success rate as proline rich binding regions, such as SH2 and SH3 domain binding sites, or hydrophobic sites, such as the MDM2 binding region of p53. Furthermore, the goodness of the prediction is also independent of the conformation the binding region adopts in the bound conformation. As most of the disordered binding regions tend to bind in either helical or coil conformation, the exclusion of either would seriously impart the usefulness of such predictor. This independency also shows the generality of ANCHOR.

The results obtained with IUPred and ANCHOR are demonstrated through the example of Human calcium/calmodulin-dependent protein kinase IV (UniProt ID: Q98TZ2), shown on Figure 1. The plot was generated with the online version of ANCHOR [104], available at http://anchor.enzim.hu/. Calcium/calmodulin-dependent kinase IV binds to calmodulin near its C-terminal end (residues 327–346). This patch is correctly identified using ANCHOR as shown in the figure. The binding region can also be identified based on one of the subclasses of calmodulin binding motifs, namely the basic 1-8-14 binding motif ([RK][RK][RK][FILVW]......[FAILVW]......[FILVW]). The location of this motif is also indicated on the Figure 1.

Despite the similar underlying philosophy of IUPred and ANCHOR, their prediction profiles are quite independent of each other. Generally, disordered binding regions can appear within highly disordered segments as well as more ordered regions. In the case of PONDR VL-XT, however, it was noted, that disordered binding regions are often

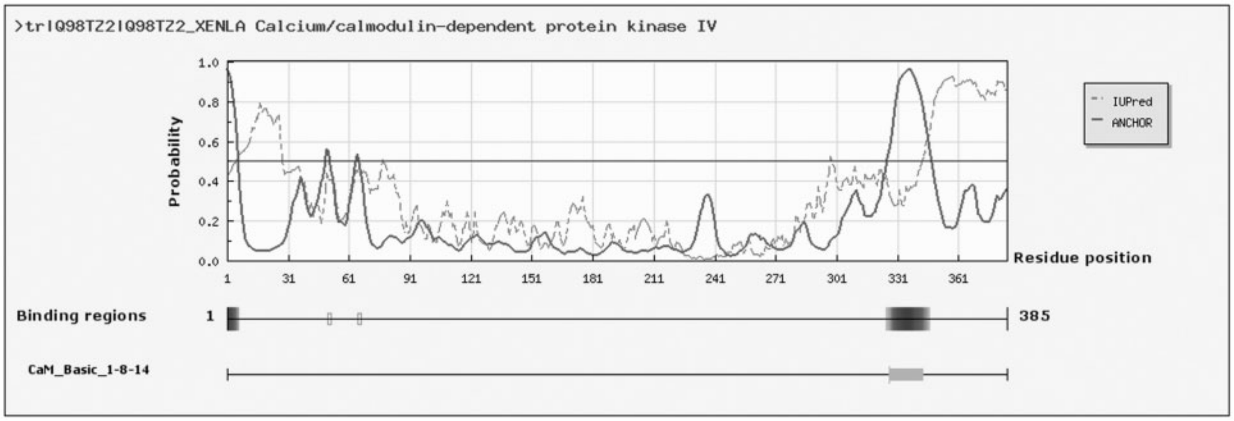

Figure 1: Output of the ANCHOR prediction server for calcium/calmodulin-dependent protein kinase IV. The plot shows the predicted disordered binding regions with the output of the general prediction method IUPred and the location of the calmodulin binding motif.

indicated by a locally more ordered region within a longer segment predicted as generally disordered [105]. Based on this finding, a specialized method was developed to recognize regions adopting α-helical conformation in their bound state, termed α-MoRFs [84]. The α-MoRF-PredII predictor employs two steps of prediction; first potential binding regions are identified by short, well pronounced dips in the VL-XT prediction score. Then these potential regions are filtered using a neural network that was trained on a selection of known α-helical disordered binding sites using sequence features such as disorder, secondary-structure predictions and amino acid indices [84, 85].

Disorder predictions can help to improve the recognition of the binding partner of certain proteins as well. For example, disorder is common within the binding partners of calmodulin, and incorporating information about predicted disorder can greatly improve the identification of potential binding partners [106]. Disorder also plays an important role in protein phosphorylation [12]. A recent method, DISPHOS [29], combines position specific amino acid frequencies with disorder information and achieves a better discrimination between phosphorylated and non-phosphorylated sites. Overall, these specific predictors can recognize subsets of disordered proteins and can provide information about their potential function.

EXAMPLES

As the case of disordered binding regions indicates, the strict categorization into order and disorder is a great oversimplification. Disorder is a complex phenomenon, and there are many examples that go beyond the classical ordered/disordered classes. In these cases, there is no single good answer from the perspective of predictions. The inability of prediction methods to handle various types of protein disorder causes a serious limitation in their efficiency. We illustrate this problem through three examples that contain a coiled-coil, a molten globule, or a disordered binding region. Although none of them have a stable 3D structure on their own, they exhibit strong structural preferences. This places them at the borderline of order and disorder in various aspects. The comparison of the behavior of several disordered prediction methods can provide insights into their general features and usability.

Heat shock factor-binding protein 1 (HSBP1)

Human HSBP1 consists of 76 amino acids. It can bind to the heat shock 70 kDa protein (HSP70) and the heat shock factor protein 1 (HSF1). Via these interactions HSBP1 negatively regulates the heat shock response [107]. HSBP1 does not have a stable structure on its own in monomeric form; however it naturally forms a homotrimer. The approximate regions between residues 8–58 of the three chains interact and form a coiled-coil structure with the rest of the protein remaining disordered [108].

This transient structure causes all of the tested prediction algorithms to react with a significantly lower score on the coiled-coil region than on the disordered N- and C-terminal parts, albeit to very different extents. VSL2B, VL3 and DisPSSMP predict the coiled-coil region to be rather disordered, POODLE-I, IUPred and RONN give a prediction of ~0.5 reflecting its ambiguous order/disorder character, and the rest of the predictors clearly predict it to be mostly ordered. While most of the prediction scores are homogenous along the coiled-coil region, there are a few profiles that distinguish certain areas. However, no conclusion can be drawn from these peaks regarding structural properties, due to the lack of clear consensus. The 13 disorder profiles are shown on Figure 2. Methods trained on short disordered regions defined by missing residues of X-ray structures only and those that rely on secondary-structure predictions heavily (e.g. DRIP-PRED) (see Table 1) clearly emphasize the tendency of the coiled region to become ordered.

Human CREB binding protein (CBP)

CBP is a 2442 residue long nuclear protein that can acetylate histones and non-histone proteins, like the NCOA3 coactivator. Among others, it binds to phosphorylated cyclic AMP-responsive element-binding protein (pCREB) and up-regulates its transcriptional activity [109, 110]. CBP is involved in transcriptional regulation and host-virus interactions as well [111]. Its segment between 2059 and 2117 constitutes the nuclear coactivator binding domain (NCBD) that is able to bind to the ACTR domain of p160 [112]. NCBD is a part of the CREB binding region (residues 2016–2115) and in its unbound form forms a molten globule that exhibits a high degree of structural order and has a large α-helical content. However, NCBD is not a folded domain

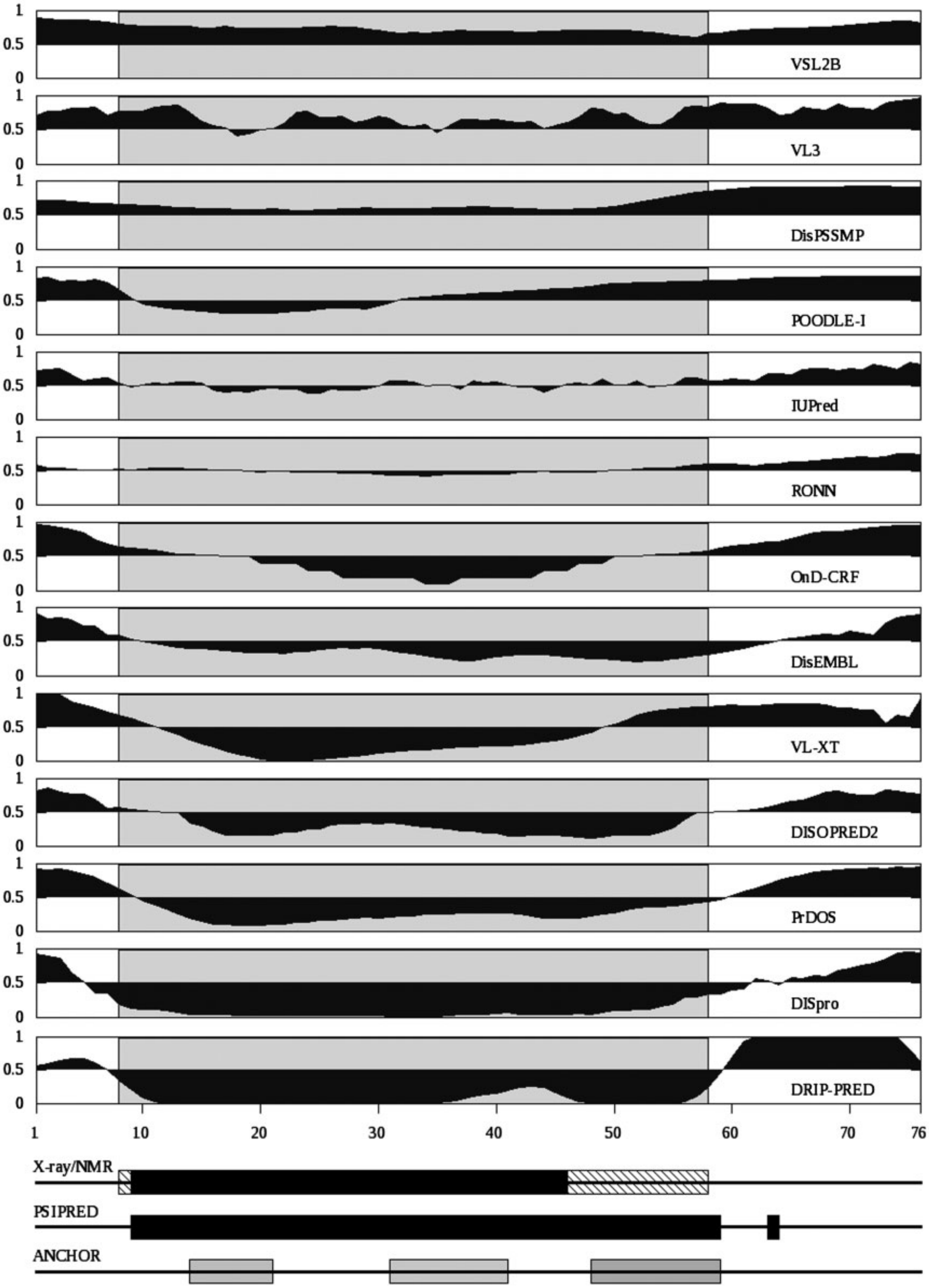

Figure 2: Disorder predictions for human HSBPI (UniProt AC: O75506). In the case of DisPSSMP, OnD-CRF and DISOPRED2, the original prediction scores were rescaled linearly to be directly comparable with other methods. Disorder predictions were sorted top to bottom by decreasing average predicted disorder tendency on the coiled coil region that is marked by the grey box on the prediction outputs. Underneath the prediction outputs, the sequence parts that were shown experimentally to adopt α-helical structure in the trimeric form either by X-ray diffraction (based on the PDB entry 3ci9 [120], marked by black box) or NMR and other experimental techniques (marked by shaded box) are shown ('X-ray/NMR'). The middle line ('PSIPRED') shows the secondary-structure prediction by PSIPRED [33], black boxes indicating predicted α-helical structure. The bottom line shows the disordered binding site prediction by ANCHOR. Shading of the boxes corresponds to the overall confidence of the predicted binding region, with black corresponding to maximal confidence.

on its own, it only adopts a well-defined tertiary structure upon binding to the fully disordered ACTR [113].

The 13 prediction profiles are shown on Figure 3 for the CREB binding domain of CBP. The order of methods according to the average predicted disordered on this proteins is remarkably similar to the one measured on HSBP1. VSL2B, VL3 and DisPSSMP predict NCBD to be generally disordered while the rest of the methods either give borderline predictions ~0.5 (as in the case of IUPred or OnD-CRF) or predict regions to be ordered. However, due to the uneven structural propensities of the protein, most of the predictions are not homogenous either. About half of the predictors respond to the more pronounced α-helical preferences of NCBD on the last two α-helices between 2084 and 2111 with strong dips in the scores. It is also clear that methods that involve secondary-structure predictions (DisPSSMP, POODLE-I, OnD-CRF, DRIP-PRED and DISpro) tend to respond to the three helices more uniformly than those that omit this kind of information. This also shows that in cases that are not unambiguously ordered or disordered, different prediction methods can behave drastically differently. These differences are not only reflected in the average amount of disorder predicted, but also in the resolution of predictions concerning underlying transient structural elements.

Human calpastatin

Calpastatin is a 708 residue long protein that is a specific inhibitor of calpain, a Ca^{2+} activated cystein protease [114]. The calpain–calpastatin interaction is part of multiple larger networks of interactions involved in the regulation of cell division, cell motility and muscle protein degradation [115]. Calpastatin contains four repeats of the calpain inhibitory domain and thus is able to inhibit four different calpain molecules at the same time. Each inhibitory domain binds to calpain via three separate binding sites (A, B and C). The center binding site B binds to the active site of calpain in an extended conformation, while the other two sites A and C bind as α-helices and increase the specificity of the interaction between the two molecules. Although calpastatin is fully disordered along its entire length, the binding sites exhibit considerable transient structure in isolated form as well [116]. These transient, preformed structural elements correspond to the secondary-structure these segments adopt upon binding to calpain, namely α-helical structure for sites A and C but site B also has highly non-random conformational preferences.

Figure 4 shows the 13 prediction profiles for the first inhibitory domain of calpastatin (137–277). While most of the predictors respond to the preformed structural elements in these sites similarly to the structural motifs present in both HSBP1 and CBP, there are a few differences. DRIP-PRED does not react to transient structure at all and simply goes into overload giving a maximal score of 1.0 through the whole domain. POODLE-I and DISOPRED2 also give scores close to 1.0, however DISOPRED2 shows some slight dips in binding sites B and C. These methods are clearly not sensitive to the capability of these binding sites to undergo a disorder-to-order transition. The rest of the predictors behave similarly to previous two examples. This shows that although some predictors tend to predict more disorder than others, this order among predictors can be heavily rearranged in the presence of different underlying propensity for structural order.

Although the dips apparent near the three binding sites are more consistent among different methods than in the previous cases, they react to these segments in a variety of ways. Some predictors only react to the general structural content of the inhibitory domain as a whole and give a slight dip in the middle of the domain coinciding with binding site B (VSL2B, DisPSSMP and POODLE-I), while some others give three distinct dips approximately corresponding to the three separate binding regions (IUPred, OnD-CRF, RONN, PrDOS, DisEMBL and DISpro). The average score on linker regions between binding sites is generally larger than the linker regions between the α-helices in CBP NCBD. This is due to the more flexible nature of these regions—as opposed to the case of CBP NCBD, these regions retain their disordered nature even in the bound form. On the other hand, the large variation in the prediction scores on the binding regions show that at these regions a naïve consensus prediction is either meaningless or very misleading.

HOW TO USE DISORDER PREDICTORS

A typical use of disorder predictions involve large-scale studies often made on whole proteomes.

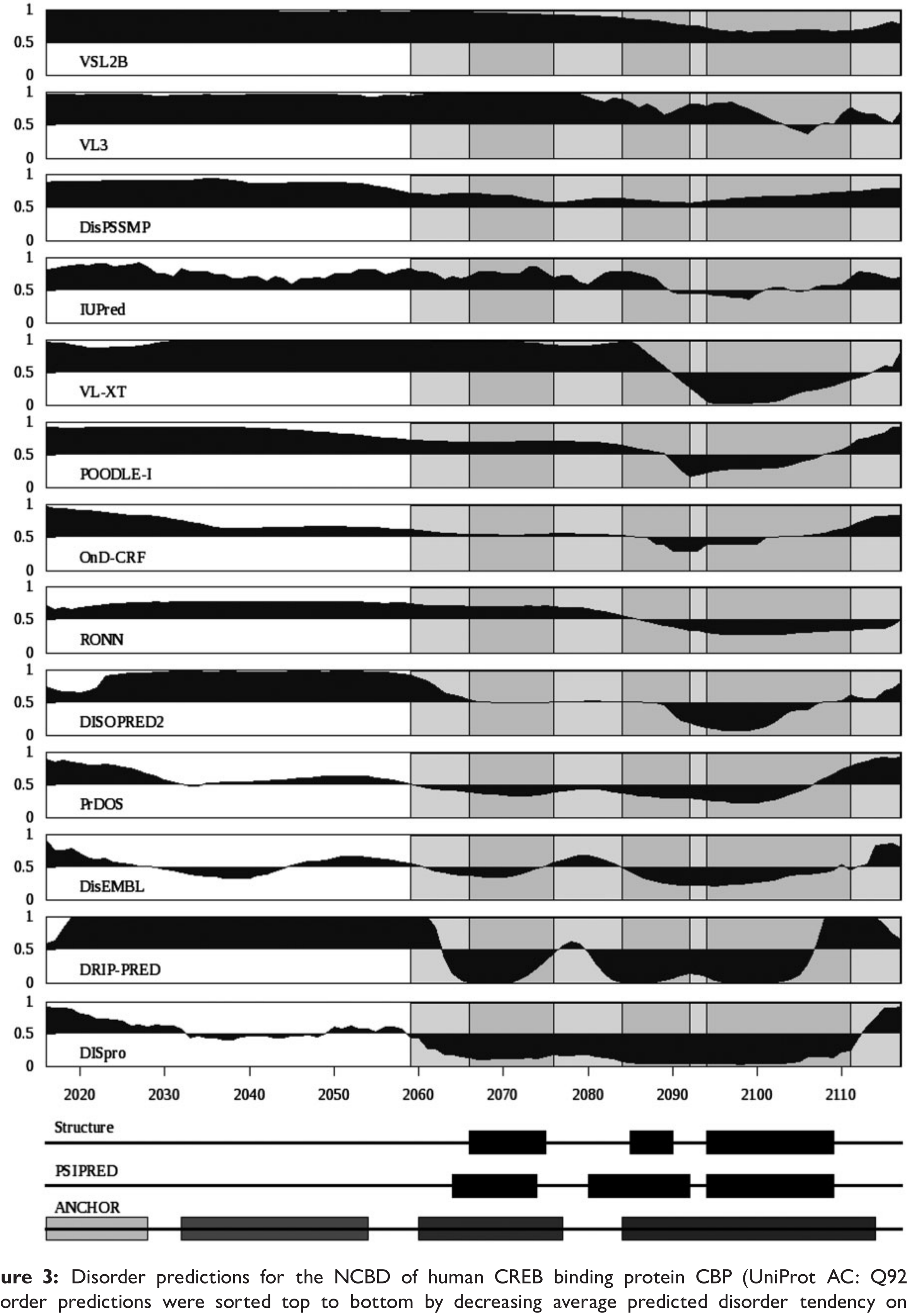

Figure 3: Disorder predictions for the NCBD of human CREB binding protein CBP (UniProt AC: Q92793). Disorder predictions were sorted top to bottom by decreasing average predicted disorder tendency on the molten globule that is marked by the light grey box on the prediction outputs, while dark grey boxes show the three α-helical regions. Underneath the prediction outputs, the sequence parts that were shown experimentally to adopt α-helical structure when bound to ACTR (based on the PDB entry 1kbh [112]) are shown ('structure'). For other details see Figure 2 legend.

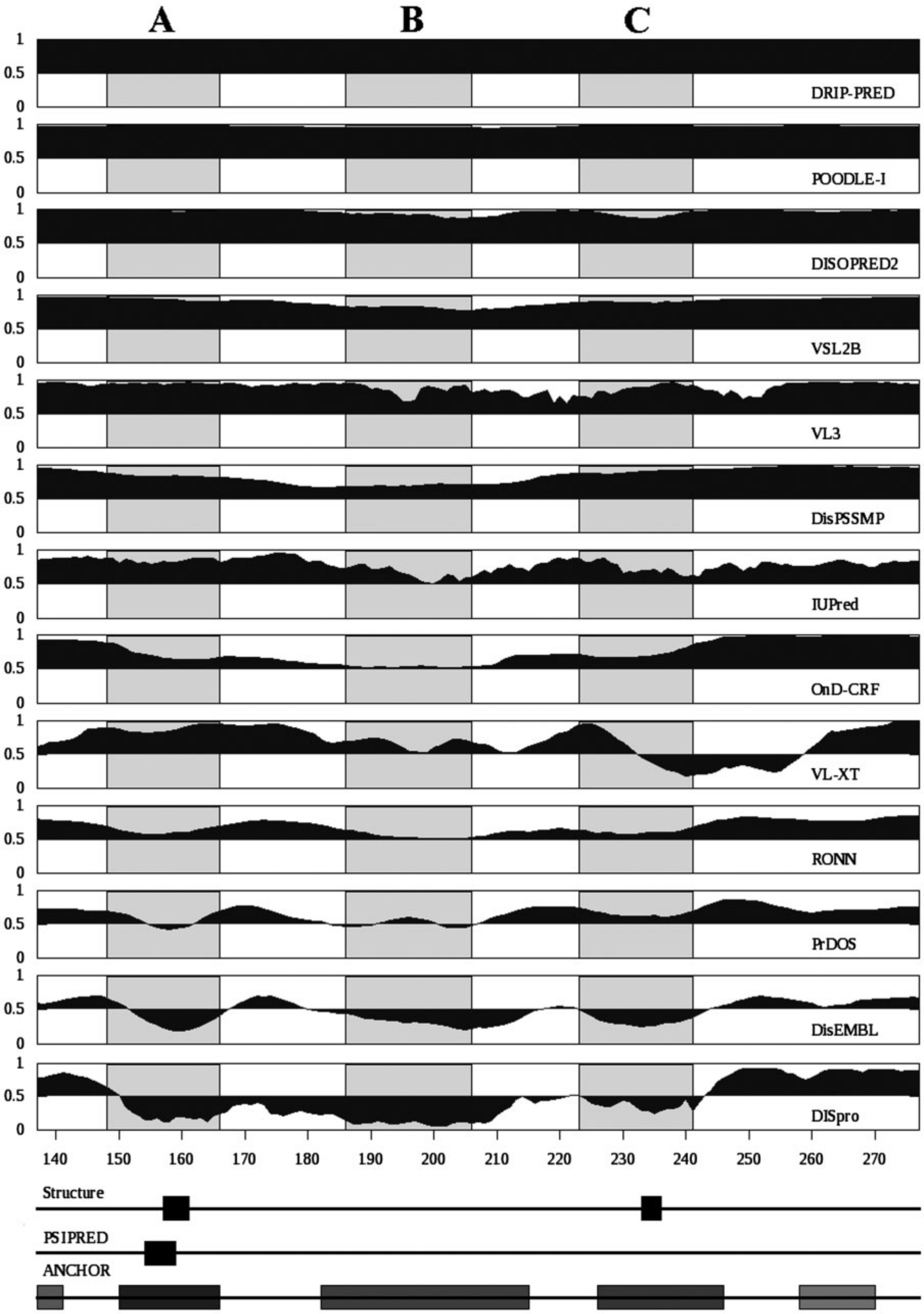

Figure 4: Disorder predictions for the first inhibitory domain of human calpastatin (UniProt AC: P20810). Disorder predictions were sorted top to bottom by decreasing average predicted disorder tendency calculated on the shown sequence part. Grey boxes labeled A, B and C on the prediction outputs mark the three binding regions. Underneath the prediction outputs, the sequence parts that were shown experimentally to adopt α-helical structure when bound to calpain (based on the PDB entry 3df0 [121]) are shown ('structure'). For other details see Figure 2 legend.

In these cases usually only longer disordered segments (typically over 30 continuous residues) are considered. Another useful term is the introduction of fully disordered proteins that can be defined by the lack of long continuous segments of predicted order (also over 30 residues). This necessitates the use of predictors that recognize larger stretches of disordered residues, such as IUPred, RONN, DisPSSMP or PONDR VL3. These methods are relatively insensitive to short regions of both order and disorder and as such usually do not segment globular or disordered domains. The output of the predictor is converted to a binary classification using a cutoff value in the prediction score. This loss of information is compensated for by the sheer volume of the analyzed sequences. It should be taken into account that two predictors can differ simply because they work at different false predictions rates. Indeed, the false prediction rate of existing methods range from 2% to nearly 15%. Calibrating the methods on the same dataset to the same false positive rate can result in more meaningful comparison.

Upon investigating single sequences, much more detail can be extracted from the prediction outputs. State-of-the-art prediction algorithms usually assign continuous scores (typically in the range of [0,1]) to residues. Again, starting by using predictors sensitive to larger regions of disorder and order, such as IUPred, RONN and PONDR VSL3, it is possible to map out globular and disordered domains. Next, predictors that can react to local disorder, such as DISpro and DISOPRED2 can be used to detect smaller disordered segments inside the globular domains. Furthermore, domain databases such as Pfam can also be incorporated into the annotation procedure that enables the distinction of flexible loops inside or between folded domains and possible checking of the location of globular domains [117]. Domain boundary predictions can also help to distinguish flexible loops from domain boundaries [118].

Another major difference between methods that should be kept in mind is their different response to partially or transiently ordered segments as shown in the above discussed three examples. On regions that are clearly fully ordered of fully disordered, most of the predictors are in agreement. However, some predictors, such as PONDR VSL2B, tend to react to transient structure only in a limited way, showing them to be mostly disordered. Other methods, such as DISpro, show these regions to be almost fully ordered. Generally, methods trained on disordered segments collected from PDB structures are generally biased towards order in these cases. The borderline characters are well-reflected in the output of IUPred or RONN, which gives predictions ~0.5 on these segments.

Dedicated predictions can reveal coiled-coil regions, while secondary-structure predictors can predict secondary-structural elements adopted in the bound state. Isolated, partially ordered α-helices can be predicted with the AGADIR method [119]. It is known that PONDR VL-XT usually responds to disordered binding sites with α-helical structural propensities with sharp dips in the prediction score. Furthermore, binding regions can be detected by using ANCHOR, regardless of secondary-structural preferences. On the other hand, larger regions that appear as ordered or contain multiple, clustered segments of local order in some predictions without a clear consensus between the output of different algorithms can be a signature of molten globules. Generally it is a good idea not to rely on one single algorithm when annotating unknown sequences. Instead, as these algorithms all capture different aspects of the structural properties of proteins, in certain cases they can complement each other to give a more complete picture.

CONCLUSION

The importance of intrinsically unstructured/disordered proteins has been recognized relatively recently as a result of large-scale genome projects and advances in experimental techniques. These proteins exist as highly flexible structural ensembles, yet they carry out vital functions in living cells and are often involved in signaling and regulatory processes. Their specific functional modes are directly linked to their intrinsic flexibility. However, this also makes them challenging subjects for experimental studies. Therefore, bioinformatic tools are indispensible for their characterization. In the last decade, several bioinformatic tools have been developed to study these proteins and some of their properties. However, disordered proteins are quite heterogeneous and existing methods can capture this only partially. This suggests that approaches that go beyond the binary classification of proteins as ordered or disordered are necessary. Recently, it was shown that the phenomenon of the lack of structure can be understood on the basis of the energy of interresidue interactions. Using this concept, not only disordered segments, but regions undergoing disorder-to-order transition could be recognized as well. This suggests that simple models incorporating basic biophysical properties of disordered segments hold the key to more detailed predictions of protein disorder.

Key Points

- Intrinsically disordered/unstructured proteins exist without a well-defined structure but carry out vital functions in the cell.
- Intrinsically unstructured proteins can be recognized from the amino acid sequence by various machine-learning algorithms.
- These proteins cannot adopt a well-defined structure because of their amino acid sequence that does not allow the formation of enough favorable interactions. This property is directly used in the IUPred method to recognize disordered proteins. Based on similar principles, ANCHOR predicts disordered binding regions.
- Protein disorder is a heterogeneous phenomenon. Methods that can target various types of disorder are needed.

FUNDING

Hungarian Scientific Research Fund (OTKA) [grant number K 72569] and National Office for Research and Technology (NKTH) [NKTH07a-TB-INTER].

Acknowledgements

The authors would like to thank Mark Adamsbaum for his critical comments on the manuscript. The fruitful discussions with Dr Peter Tompa are gratefully acknowledged.

References

1. Dutta S, Burkhardt K, Young J, *et al.* Data deposition and annotation at the worldwide protein data bank. *Mol Biotechnol* 2009;**42**:1–13.
2. Dyson HJ, Wright PE. Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol* 2005;**6**:197–208.
3. Wright PE, Dyson HJ. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J Mol Biol* 1999;**293**:321–31.
4. Dunker AK, Lawson JD, Brown CJ, *et al.* Intrinsically disordered protein. *J Mol Graph Model* 2001;**19**:26–59.
5. Tompa P. Intrinsically unstructured proteins. *Trends Biochem Sci* 2002;**27**:527–533.
6. Uversky VN, Gillespie JR, Fink AL. Why are ‘natively unfolded’ proteins unstructured under physiologic conditions? *Proteins* 2000;**41**:415–27.
7. Dunker AK, Obradovic Z, Romero P, *et al.* Intrinsic protein disorder in complete genomes. *Genome Inform Ser Workshop Genome Inform* 2000;**11**:161–71.
8. Ward JJ, Sodhi JS, McGuffin LJ, *et al.* Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol* 2004;**337**:635–45.
9. Liu J, Perumal NB, Oldfield CJ, *et al.* Intrinsic disorder in transcription factors. *Biochemistry* 2006;**45**:6873–88.
10. Fuxreiter M, Tompa P, Simon I, *et al.* Malleable machines take shape in eukaryotic transcriptional regulation. *Nat Chem Biol* 2008;**4**:728–37.
11. 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**:1882–98.
12. Galea CA, Wang Y, Sivakolundu SG, *et al.* Regulation of cell division by intrinsically unstructured proteins: intrinsic flexibility, modularity, and signaling conduits. *Biochemistry* 2008;**47**:7598–609.
13. Dunker AK, Cortese MS, Romero P, *et al.* Flexible nets. The roles of intrinsic disorder in protein interaction networks. *Febs J* 2005;**272**:5129–48.
14. Uversky VN, Oldfield CJ, Dunker AK. Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling. *J Mol Recognit* 2005;**18**:343–84.
15. Iakoucheva LM, Brown CJ, Lawson JD, *et al.* Intrinsic disorder in cell-signaling and cancer-associated proteins. *J Mol Biol* 2002;**323**:573–84.
16. Uversky VN, Oldfield CJ, Dunker AK. Intrinsically disordered proteins in human diseases: introducing the D2 concept. *Annu Rev Biophys* 2008;**37**:215–46.
17. Cheng Y, LeGall T, Oldfield CJ, *et al.* Rational drug design via intrinsically disordered protein. *Trends Biotechnol* 2006;**24**: 435–42.
18. Uversky VN. Natively unfolded proteins: a point where biology waits for physics. *Protein Sci* 2002;**11**:739–56.
19. McCarney ER, Werner JH, Bernstein SL, *et al.* Site-specific dimensions across a highly denatured protein; a single molecule study. *J Mol Biol* 2005;**352**:672–82.
20. Dosztányi Z, Sándor M, Tompa P, *et al.* Prediction of protein disorder at the domain level. *Curr Protein Pept Sci* 2007; **8**:161–71.
21. Le Gall T, Romero PR, Cortese MS, *et al.* Intrinsic disorder in the Protein Data Bank. *J Biomol Struct Dyn* 2007;**24**: 325–42.
22. Tompa P. The interplay between structure and function in intrinsically unstructured proteins. *FEBS Lett* 2005;**579**: 3346–54.
23. Dyson HJ, Wright PE. Coupling of folding and binding for unstructured proteins. *Curr Opin Struct Biol* 2002;**12**: 54–60.
24. Dunker AK, Garner E, Guilliot S, *et al.* Protein disorder and the evolution of molecular recognition: theory, predictions and observations. *Pac Symp Biocomput* 1998;473–84.
25. Gunasekaran 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**:1327–41.
26. Mészáros B, Tompa P, Simon I, *et al.* Molecular principles of the interactions of disordered proteins. *J Mol Biol* 2007; **372**:549–61.
27. Vacic V, Oldfield CJ, Mohan A, *et al.* Characterization of molecular recognition features, MoRFs, and their binding partners. *J Proteome Res* 2007;**6**:2351–66.

28. 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**(Suppl 1):S1.
29. Iakoucheva LM, Radivojac P, Brown CJ, *et al*. The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res* 2004;**32**:1037–49.
30. Bracken C, Iakoucheva LM, Romero PR, *et al*. Combining prediction, computation and experiment for the characterization of protein disorder. *Curr Opin Struct Biol* 2004;**14**: 570–6.
31. Receveur-Brechot V, Bourhis JM, Uversky VN, *et al*. Assessing protein disorder and induced folding. *Proteins* 2006;**62**:24–45.
32. Rost B, Sander C. Prediction of protein secondary structure at better than 70% accuracy. *J Mol Biol* 1993;**232**: 584–99.
33. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* 1999;**292**: 195–202.
34. Cuff JA, Barton GJ. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins* 2000;**40**:502–11.
35. Rost B, Sander C. Conservation and prediction of solvent accessibility in protein families. *Proteins* 1994;**20**:216–26.
36. Adamczak R, Porollo A, Meller J. Accurate prediction of solvent accessibility using neural networks-based regression. *Proteins* 2004;**56**:753–67.
37. Ceroni A, Passerini A, Vullo A, *et al*. DISULFIND: a disulfide bonding state and cysteine connectivity prediction server. *Nucleic Acids Res* 2006;**34**:W177–81.
38. Lippi M, Passerini A, Punta M, *et al*. MetalDetector: a web server for predicting metal-binding sites and disulfide bridges in proteins from sequence. *Bioinformatics* 2008;**24**: 2094–5.
39. Fiser A, Simon I. Predicting redox state of cysteines in proteins. *Methods Enzymol* 2002;**353**:10–21.
40. Kim DE, Chivian D, Malmstrom L, *et al*. Automated prediction of domain boundaries in CASP6 targets using Ginzu and RosettaDOM. *Proteins* 2005;**61**(Suppl 7):193–200.
41. Yoo PD, Sikder AR, Zhou BB, *et al*. Improved general regression network for protein domain boundary prediction. *BMC Bioinformatics* 2008;**9**(Suppl 1):S12.
42. Bryson K, Cozzetto D, Jones DT. Computer-assisted protein domain boundary prediction using the DomPred server. *Curr Protein Pept Sci* 2007;**8**:181–8.
43. Ofran Y, Rost B. ISIS: interaction sites identified from sequence. *Bioinformatics* 2007;**23**:e13–6.
44. Tusnády GE, Simon I. Principles governing amino acid composition of integral membrane proteins: application to topology prediction. *J Mol Biol* 1998;**283**:489–506.
45. Sonnhammer EL, von Heijne G, Krogh A. A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc Int Conf Intell Syst Mol Biol* 1998;**6**: 175–82.
46. Cserző M, Eisenhaber F, Eisenhaber B, *et al*. TM or not TM: transmembrane protein prediction with low false positive rate using DAS-TMfilter. *Bioinformatics* 2004;**20**:136–7.
47. Ferron F, Longhi S, Canard B, *et al*. A practical overview of protein disorder prediction methods. *Proteins* 2006;**65**:1–14.
48. Dosztányi Z, Tompa P. Prediction of protein disorder. *Methods Mol Biol* 2008;**426**:103–15.
49. He B, Wang K, Liu Y, *et al*. Predicting intrinsic disorder in proteins: an overview. *Cell Res* 2009;**19**:929–49.
50. Berman HM, Westbrook J, Feng Z, *et al*. The Protein Data Bank. *Nucleic Acids Res* 2000;**28**:235–42.
51. Sickmeier M, Hamilton JA, LeGall T, *et al*. DisProt: the database of disordered proteins. *Nucleic Acids Res* 2007;**35**: D786–93.
52. Shimizu K, Muraoka Y, Hirose S, *et al*. Predicting mostly disordered proteins by using structure-unknown protein data. *BMC Bioinformatics* 2007;**8**:78.
53. Mohan A, Uversky VN, Radivojac P. Influence of sequence changes and environment on intrinsically disordered proteins. *PLoS Comput Biol* 2009;**5**:e1000497.
54. De Biasio A, Guarnaccia C, Popovic M, *et al*. Prevalence of intrinsic disorder in the intracellular region of human single-pass type I proteins: the case of the notch ligand Delta-4. *J Proteome Res* 2008;**7**:2496–506.
55. Melamud E, Moult J. Evaluation of disorder predictions in CASP5. *Proteins* 2003;**53**(Suppl 6):561–5.
56. Jin Y, Dunbrack RL, Jr. Assessment of disorder predictions in CASP6. *Proteins* 2005;**61**(Suppl 7):167–175.
57. Bordoli L, Kiefer F, Schwede T. Assessment of disorder predictions in CASP7. *Proteins* 2007;**69**(Suppl 8): 129–36.
58. Noivirt-Brik O, Prilusky J, Sussman JL. Assessment of disorder predictions in CASP8. *Proteins* 2009;**77**(Suppl 9): 210–6.
59. Peng K, Radivojac P, Vucetic S, *et al*. Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics* 2006;**7**:208.
60. Schlessinger A, Punta M, Yachdav G, *et al*. Improved disorder prediction by combination of orthogonal approaches. *PLoS ONE* 2009;**4**:e4433.
61. Hirose S, Shimizu K, Kanai S, *et al*. POODLE-L: a two-level SVM prediction system for reliably predicting long disordered regions. *Bioinformatics* 2007;**23**:2046–53.
62. Garner E, Cannon P, Romero P, *et al*. Predicting disordered regions from amino acid sequence: common themes despite differing structural characterization. *Genome Inform Ser Workshop Genome Inform* 1998;**9**:201–13.
63. Radivojac P, Obradovic Z, Smith DK, *et al*. Protein flexibility and intrinsic disorder. *Protein Sci* 2004;**13**:71–80.
64. Li X, Romero P, Rani M, *et al*. Predicting protein disorder for N-, C-, and internal regions. *Genome Inform Ser Workshop Genome Inform* 1999;**10**:30–40.
65. Xie Q, Arnold GE, Romero P, *et al*. The sequence attribute method for determining relationships between sequence and protein disorder. *Genome Inform Ser Workshop Genome Inform* 1998;**9**:193–200.
66. Campen A, Williams RM, Brown CJ, *et al*. TOP-IDP-scale: a new amino acid scale measuring propensity for intrinsic disorder. *Protein Pept Lett* 2008;**15**:956–63.
67. Galzitskaya OV, Garbuzynskiy SO, Lobanov MY. FoldUnfold: web server for the prediction of disordered regions in protein chain. *Bioinformatics* 2006;**22**:2948–9.
68. Linding R, Russell RB, Neduva V, *et al*. GlobPlot: exploring protein sequences for globularity and disorder. *Nucleic Acids Res* 2003;**31**:3701–8.
69. Su CT, Chen CY, Ou YY. Protein disorder prediction by condensed PSSM considering propensity for order or disorder. *BMC Bioinformatics* 2006;**7**:319.

70. Wootton JC. Non-globular domains in protein sequences: automated segmentation using complexity measures. *Comput Chem* 1994;**18**:269–85.

71. Wootton JC, Federhen S. Analysis of compositionally biased regions in sequence databases. *Methods Enzymol* 1996;**266**:554–71.

72. Romero P, Obradovic Z, Li X, *et al*. Sequence complexity of disordered protein. *Proteins* 2001;**42**:38–48.

73. Dosztányi Z, Chen J, Dunker AK, *et al*. Disorder and sequence repeats in hub proteins and their implications for network evolution. *J Proteome Res* 2006;**5**:2985–95.

74. Schaffer AA, Aravind L, Madden TL, *et al*. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res* 2001;**29**:2994–3005.

75. Brown CJ, Takayama S, Campen AM, *et al*. Evolutionary rate heterogeneity in proteins with long disordered regions. *J Mol Evol* 2002;**55**:104–10.

76. Daughdrill GW, Narayanaswami P, Gilmore SH, *et al*. Dynamic behavior of an intrinsically unstructured linker domain is conserved in the face of negligible amino acid sequence conservation. *J Mol Evol* 2007;**65**:277–88.

77. Fuxreiter M, Simon I, Friedrich P, *et al*. Preformed structural elements feature in partner recognition by intrinsically unstructured proteins. *J Mol Biol* 2004;**338**:1015–26.

78. Süveges D, Gáspári Z, Tóth G, *et al*. Charged single alpha-helix: a versatile protein structural motif. *Proteins* 2009;**74**: 905–16.

79. Cheng J, Sweredoski M, Baldi P. Accurate prediction of protein disordered regions by mining protein structure data. *Data Mining and Knowledge Discovery* 2005;**11**:213–22.

80. Su CT, Chen CY, Hsu CM. iPDA: integrated protein disorder analyzer. *Nucleic Acids Res* 2007;**35**:W465–72.

81. Wang L, Sauer UH. OnD-CRF: predicting order and disorder in proteins using [corrected] conditional random fields. *Bioinformatics* 2008;**24**:1401–2.

82. Schlessinger A, Liu J, Rost B. Natively unstructured loops differ from other loops. *PLoS Comput Biol* 2007;**3**:e140.

83. Romero, Obradovic, Dunker K. Sequence data analysis for long disordered regions prediction in the calcineurin family. *Genome Inform Ser Workshop Genome Inform* 1997;**8**: 110–24.

84. Cheng Y, Oldfield CJ, Meng J, *et al*. Mining alpha-helix-forming molecular recognition features with cross species sequence alignments. *Biochemistry* 2007;**46**:13468–77.

85. Oldfield CJ, Cheng Y, Cortese MS, *et al*. Coupled folding and binding with alpha-helix-forming molecular recognition elements. *Biochemistry* 2005;**44**:12454–70.

86. Radivojac P, Obradovic Z, Brown CJ, *et al*. Prediction of boundaries between intrinsically ordered and disordered protein regions. *Pac Symp Biocomput* 2003;216–27.

87. Linding R, Jensen LJ, Diella F, *et al*. Protein disorder prediction: implications for structural proteomics. *Structure* 2003;**11**:1453–9.

88. Yang ZR, Thomson R, McNeil P, *et al*. RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics* 2005;**21**:3369–76.

89. Ishida T, Kinoshita K. PrDOS: prediction of disordered protein regions from amino acid sequence. *Nucleic Acids Res* 2007;**35**:W460–4.

90. Shimizu K, Hirose S, Noguchi T. POODLE-S: web application for predicting protein disorder by using physicochemical features and reduced amino acid set of a position-specific scoring matrix. *Bioinformatics* 2007;**23**:2337–8.

91. MacCallum RM.2004. http://www.forcasp.org/paper2127.html (11 November 2009, date last accessed).

92. Xue B, Oldfield CJ, Dunker AK, *et al*. CDF it all: consensus prediction of intrinsically disordered proteins based on various cumulative distribution functions. *FEBS Lett* 2009;**583**: 1469–74.

93. Obradovic Z, Peng K, Vucetic S, *et al*. Exploiting heterogeneous sequence properties improves prediction of protein disorder. *Proteins* 2005;**61**(Suppl 7):176–82.

94. Hirose S, Shimizu K, Inoue N, *et al*. Disordered region prediction by integrating POODLE series. *Eighth Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction (CASP8) 2008*; 14–15.

95. Bujnicki JM, Elofsson A, Fischer D, *et al*. LiveBench-2: large-scale automated evaluation of protein structure prediction servers. *Proteins* 2001;**45**(Suppl 5):184–91.

96. McGuffin LJ. Intrinsic disorder prediction from the analysis of multiple protein fold recognition models. *Bioinformatics* 2008;**24**:1798–804.

97. Ishida T, Kinoshita K. Prediction of disordered regions in proteins based on the meta approach. *Bioinformatics* 2008;**24**: 1344–8.

98. Kozlowski LP, Bujnicki JM. *Eighth Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction (CASP8) 2008;* 46.

99. Prilusky J, Felder CE, Zeev-Ben-Mordehai T, *et al*. FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics* 2005;**21**: 3435–38.

100. Dosztányi Z, Csizmók V, Tompa P, *et al*. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J Mol Biol* 2005;**347**:827–39.

101. Dosztányi Z, Csizmók V, Tompa P, *et al*. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* 2005;**21**:3433–4.

102. Thomas PD, Dill KA. An iterative method for extracting energy-like quantities from protein structures. *Proc Natl Acad Sci USA* 1996;**93**:11628–33.

103. Mészáros B, Simon I, Dosztányi Z. Prediction of protein binding regions in disordered proteins. *PLoS Comput Biol* 2009;**5**:e1000376.

104. Dosztányi Z, Mészáros B, Simon I. ANCHOR: web server for predicting protein binding regions in disordered proteins. *Bioinformatics* 2009;**25**:2745–6.

105. Garner E, Romero P, Dunker AK, *et al*. Predicting binding regions within disordered proteins. *Genome Inform Ser Workshop Genome Inform* 1999;**10**:41–50.

106. 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.

107. Satyal SH, Chen D, Fox SG, *et al*. Negative regulation of the heat shock transcriptional response by HSBP1. *Genes Dev* 1998;**12**:1962–74.

108. Tai LJ, McFall SM, Huang K, *et al*. Structure-function analysis of the heat shock factor-binding protein reveals a

protein composed solely of a highly conserved and dynamic coiled-coil trimerization domain. *J Biol Chem* 2002;**277**: 735–45.

109. Goodman RH, Smolik S. CBP/p300 in cell growth, transformation, and development. *Genes Dev* 2000;**14**:1553–77.

110. Janknecht R. The versatile functions of the transcriptional coactivators p300 and CBP and their roles in disease. *Histol Histopathol* 2002;**17**:657–68.

111. Revilla Y, Granja AG. Viral mechanisms involved in the transcriptional CBP/p300 regulation of inflammatory and immune responses. *Crit Rev Immunol* 2009;**29**:131–54.

112. Demarest SJ, Martinez-Yamout M, Chung J, *et al*. Mutual synergistic folding in recruitment of CBP/p300 by p160 nuclear receptor coactivators. *Nature* 2002;**415**:549–53.

113. Demarest SJ, Deechongkit S, Dyson HJ, *et al*. Packing, specificity, and mutability at the binding interface between the p160 coactivator and CREB-binding protein. *Protein Sci* 2004;**13**:203–10.

114. Goll DE, Thompson VF, Li H, *et al*. The calpain system. *Physiol Rev* 2003;**83**:731–801.

115. Wendt A, Thompson VF, Goll DE. Interaction of calpastatin with calpain: a review. *Biol Chem* 2004;**385**:465–72.

116. Kiss R, Kovács D, Tompa P, *et al*. Local structural preferences of calpastatin, the intrinsically unstructured protein inhibitor of calpain. *Biochemistry* 2008;**47**:6936–45.

117. Finn RD, Tate J, Mistry J, *et al*. The Pfam protein families database. *Nucleic Acids Res* 2008;**36**:D281–88.

118. Holland TA, Veretnik S, Shindyalov IN, *et al*. Partitioning protein structures into domains: why is it so difficult? *J Mol Biol* 2006;**361**:562–90.

119. Lacroix E, Viguera AR, Serrano L. Elucidating the folding problem of alpha-helices: local motifs, long-range electrostatics, ionic-strength dependence and prediction of NMR parameters. *J Mol Biol* 1998;**284**:173–91.

120. Liu X, Xu L, Liu Y, *et al*. Crystal structure of the hexamer of human heat shock factor binding protein 1. *Proteins* 2009; **75**:1–11.

121. Moldoveanu T, Gehring K, Green DR. Concerted multipronged attack by calpastatin to occlude the catalytic cleft of heterodimeric calpains. *Nature* 2008;**456**:404–8.