Bioinformatical Approaches to Unstructured/ Disordered Proteins and Their Interactions

Bálint Mészáros, Zsuzsanna Dosztányi, Csaba Magyar, and István Simon

Abstract. Intrinsically unstructured/disordered proteins (IUPs/IDPs) exist as highly flexible conformational ensembles without adopting a stable three-dimensional structure. Experimental and bioinformatical studies in the past two decades have shown that these proteins play a central role in various signaling and regulatory processes. Accordingly, their frequency in higher eukaryotes reaches high proportions and their malfunction can be connected to a wide variety of diseases. Recognizing the biological importance of these proteins motivated researchers to understand various aspects of disordered proteins and protein segments from the viewpoint of biochemistry, molecular biology and pharmacology. In general, IDPs are difficult to study experimentally because of the lack of a unique structure in the isolated form. Nevertheless, various bioinformatics tools developed over the last few years enable their identification and characterization using only the amino acid sequence. In this chapter — after a brief introduction to IDPs in general — we present a small survey of current methods aimed at identifying disordered proteins or protein segments, focusing on those that are publicly available as web servers. We also discuss in more detail approaches that predict disordered regions and specific regions involved in protein binding by modeling the physical background of protein disorder. Furthermore, we argue that the heterogeneity of disordered segments needs to be taken into account for a better understanding of protein disorder and the correct use and interpretation of the output of disorder prediction algorithms.

Bálint Mészáros · Zsuzsanna Dosztányi · Csaba Magyar · István Simon
Institute of Enzymology, RCNS, HAS; Budapest, Hungary
e-mail: {meszaros.balint,dosztanyi.zsuzsanna,
magyar.csaba,simon.istvan}@ttk.mta.hu

A. Liwo (Ed.): *Comput. Methods to Study the Struct. & Dyn. of Biomolecules*, SSBN 1, pp. 525–556.
DOI: 10.1007/978-3-642-28554-7_16

1 Introduction to Disordered Proteins

In approximately the first 40 years of structural biology, the central model underlying all biochemical studies was that a well-formed structure is a prerequisite for a protein to carry out its function. This notion motivated a large number of structure-function studies and led to the structure determination of around 80,000 proteins as of date. Although some proteins and protein segments were known that either did not lend themselves to structure determination or had sequence features that were seemingly incompatible with a folded structure (e.g., highly charged, repetitive sequence regions), these were considered as hallmarks of imperfect experimental conditions or some exotic rarities of nature.

1.1 Re-assessing the Structure-Function Paradigm

With the explosion of available genome sequences, during the 1990s the known number of these 'rarities' and 'experimental errors' grew steadily to the point where they could no longer be written down on a side note. This forced molecular biologists to reassess the structure-function paradigm[1]. The world of proteins was extended to include proteins that do not require a stable, three dimensional structure even under physiological conditions in order to fulfill their biological role[2-4]. These intrinsically unstructured/disordered proteins (IUPs/IDPs) lack a well-defined tertiary structure and exhibit a multitude of conformations that dynamically change over time and population. The importance of protein disorder is underlined by the abundance of partially or fully disordered proteins encoded in higher eukaryotic genomes[5]. Using bioinformatics methods (discussed in later sections) it was estimated that 30–50% of eukaryotic proteins contain at least one long disordered segment. The fact that protein disorder is not a tolerated necessity but provides an evolutionary advantage is reflected by studies showing the steady increase of the percentage of disordered proteins in proteomes as organism complexity increases[6,7]. Furthermore, disordered proteins are involved in many critical processes[3] such as transcription, translation, regulation, signal transduction and stress-response, complementing the functional repertoire of globular proteins[8].

Recent characterization of IDPs based on their functions shows that disorder can help these proteins to fulfill their functions in various ways[9,10]. In accord with the wide variety of functions associated with it, protein disorder also comes in a variety. In some cases, disordered regions are short and can be found at the terminal regions of globular domains, such as the disordered N-terminal region of the eIF4E protein. Similarly, globular domains can also harbor flexible loops that appear as missing regions in solved structures. Flexible linkers that connect globular domains, such as zinc fingers, represent another type of localized disorder. In another scenario, especially in complex organisms, protein disorder often encompasses larger, domain sized regions. These regions can exhibit different degrees of flexibility ranging from the near-random conformation of the ACTR domain of the p160 protein, through the presence of local transient secondary structural elements — such as in the N-terminal region of p27 —, to compact molten globule regions with considerable amount of secondary structure but without stable tertiary structure, such as the nuclear coactivator binding domain of the CBP protein.

Given the functional importance of disordered protein regions, their malfunction is expected to have serious biological consequences. IDPs have been implicated in various diseases, including neurodegenerative diseases, amyloidosis, diabetes, cardiovascular diseases and cancer[11-14]. Despite the fact that proteins involved in these diseases are shown to have a higher disorder content, the exact role of protein disorder in the diseases themselves are not fully understood. Probably, most results published to date concern the involvement of IDPs in cancer[15]. BRCA1, p27, p21 and CBP are examples of proteins with a significant amount of disorder that have been associated with various forms of cancer. One of the best characterized disordered proteins, p53, is known to be directly inactivated in more than 50% of cancers. At a more general level, the higher proportion of disordered proteins among cancer associated proteins was also observed[15]. However, it has been shown that the link between protein disorder and the involvement in cancer is not casual. In fact, both are strongly correlated with protein function which links them together[16]. This clearly calls for a more detailed understanding of the role of protein disorder in various diseases.

Apart from basic research interests, the connection between protein disorder and its role in diseases has implications in therapeutics as well. The pharmaceutical industry is currently struggling to find promising new drug targets, despite substantial increases in research funding. Drug discovery rates seem to have reached a plateau or are perhaps even declining, suggesting the need for new strategies. Until recently, the feasibility of targeting proteins without a well-defined structure was unclear for the purpose of drug development[17]. There is now, however, a newly sparked interest in IDPs as potential drug targets[18]. This is supported by the finding of specific inhibitors to block the interaction between a disordered region of p53 and the folded MDM2, or between c-Myc and Max. Recognizing the relevance of these proteins stimulated more systematic efforts aimed at their structural characterization and determination of their mechanisms of action.

1.2 Coupled Folding and Binding of IDPs

As the above pharmaceutical examples show, the study of the interactions involving IDPs is of special interest that has relevance not only from a therapeutic viewpoint but also from a basic research perspective as well. With the exception of a few known disordered proteins, such as entropic chains (where the biological function is directly mediated by disorder, as in the case of the MAP2 projection domain, titin's PEVK domain and the nucleoporin complex), most disordered proteins function by binding specifically to other proteins, DNA or RNA. The lack of structure in the unbound form has a profound effect on both the binding process and the properties of the resulting complex[19,20]. In all cases, the flexibility of the disordered partner decreases due to the binding. As a result, usually the resulting complex lends itself to traditional structure determination, in which cases the folding is said to be coupled to binding. Although the Protein Data Bank (PDB[21]) contains significantly fewer such cases, these examples already demonstrate the definitive differences of the complexes involving disordered proteins compared to the complexes of ordered globular proteins. The structure of the complexes involving disordered proteins also shows a rigid conformation, however many of their distinct properties give away their inherent flexibility[22,23].

In most cases, disordered segments adopt a largely extended and open conformation in the complex. Probably one of the most characteristic features of disordered binding regions is that they are usually well localized in the sequence — in about 70% of the cases the interacting residues can be mapped to a single continuous region of residues. These localized interacting regions allow IDPs to have an increased modularity as different binding regions can be incorporated into the same protein without excessively increasing protein length. These binding regions can be close to each other or can form mutually exclusive overlapping sites creating molecular switches.

The distinct binding mode of IDPs is also reflected in the physico-chemical nature of their interfaces. These interfaces are more hydrophobic, and the preferred interaction contacts are also significantly different compared to the more familiar globular proteins. As opposed to the large number of polar-polar interactions at globular interfaces, IDPs tend to favor hydrophobic-hydrophobic contacts with the partner protein. The increased importance of hydrophobic interactions during binding is a hallmark of the complexes involving IDPs.

Figure 1 shows a protein complex with three interactors. This complex on one hand demonstrates a typical interaction between ordered proteins and on the other hand also shows an interaction between ordered proteins and a disordered protein. The shown solved structure is of the complex between the ordered cyclinA and cyclin dependent kinase 2 (CDK2) proteins inhibited by the disordered p27 protein. The interaction between cyclinA and CDK2 plays an essential role in the control of the S and G2 phases of eukaryotic cell cycle. The specific interaction between the two proteins enables CDK2 to bind ATP due to slight structural rearrangements emerging during the binding. The interaction surface is dominated by polar and charged residues and is relatively planar. The interface is of moderate size compared to the size of the proteins with about 13–14% of the residues of both cyclin A and CDK2 visible in the structure being involved directly in the binding.

A strikingly different molecular recognition scenario is presented by the disordered p27 in the complex. The segment of p27 involved in the binding shows only little helical preferences in the unbound form. However, several regions adopt a well-defined α-helix upon binding. The group of the most strongly interacting residues of p27 is dominated by hydrophobic/aromatic residues that fit into hydrophobic clefts and grooves on the surface of the cyclinA-CDK2 complex. The structure shows that the interacting region of p27 forms a largely linear binding site in the sense that all residues of p27 interacting with the ordered partner complex are sequentially close. This enables p27 to incorporate a significantly larger fraction of its residues into the interaction, and accordingly over two-thirds of the visible residues of p27 are directly involved in the binding.

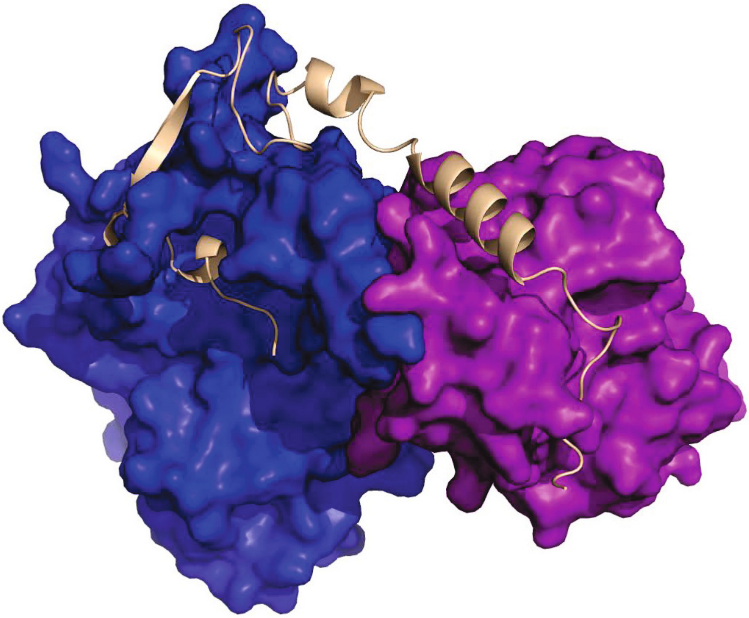


Fig. 1 Example of interfaces between two ordered proteins and a disordered protein. The ordered CDK2 and cyclinA are shown in blue and purple surface representations, respectively. The disordered p27 is shown in light salmon cartoon representation. The figure was generated from the 1jsu PDB file.

In the case of disordered proteins, the coupling between folding and binding is not only apparent in the structural properties of the resulting complex, but also in the thermodynamics and energetics of the binding. Following the basic rules of thermodynamics, the resulting protein complex corresponds to the state with an energy minimum. However — as opposed to the interaction of globular proteins — in complexes involving IDPs, the loss of entropy during the folding of the disordered partner plays a much larger role, which results in a weaker binding compared to that of globular proteins. This way the specificity, which is basically independent of the entropic terms, is uncoupled from binding strength[20]. This enables IDPs to form specific, yet transient interactions, which are indispensable to regulatory and signaling processes[3,9,10]. The increased rate of association and dissociation of disordered proteins increase their temporal binding capacity. Furthermore, disordered proteins are able to incorporate a higher fraction of their surface in the binding interface, which increases their interaction capacity in a spatial sense as well[24]. Consequently, disordered proteins in general can mediate a large number of interactions thus serving as hubs of protein-protein interaction networks[25].

1.3 Experimental Techniques and the Need for Bioinformatics

The detailed structural and functional characterization of disordered proteins is a challenging task[26]. On one hand, as disordered proteins are generally involved in regulatory functions, their expression levels are relatively low on average, making them more difficult to isolate. On the other hand, disordered regions are more prone to degradation by proteolytic enzymes than well-folded proteins. Furthermore, the existing experimental procedures are highly biased toward ordered proteins, and most techniques provide only indirect information about disorder[3]. Consequently, the current list of experimentally verified disordered proteins is rather limited with numbers in the hundreds. This is especially alarming in light of the fact that about half of the human proteins are estimated to contain at least one longer disordered segment. This discrepancy faithfully reflects the difficulties of the experimental identification of disordered proteins. Because of these difficulties, bioinformatics tools that target the prediction of protein disorder from the sequence play a very important role in the identification and characterization of IDPs as only these tools can give us information about their basic properties, evolution, and functions on a large scale.

2 Disorder Prediction Methods

As with all bioinformatics prediction algorithms, the prediction of protein disorder presents issues at several different levels. These include the buildup of the prediction algorithm itself, but the proper choice of training and testing databases and the correct evaluation of the resulting method are equally important. In the following sections, we give a brief overview of the basic concepts and techniques of disorder prediction methods.

2.1 Basic Sequence Properties of IDPs

Disordered proteins have very distinct sequence properties compared to globular proteins. These differences were already apparent when only a handful of examples for protein disorder were known. The first analyses of sequences of disordered proteins revealed that in comparison to globular proteins, these proteins are generally enriched in polar and charged amino acids at the expense of aliphatic and aromatic amino acids. At closer inspection, however, various subsets of disordered protein sequences exhibited further variations in their sequential biases. Differences in the amino acid composition could be observed depending on the experimental method used to identify disordered regions (e.g., CD, NMR or X-ray crystallography)[27], or on the location in the sequence (N- and C-terminal, middle regions)[28]. Shorter and longer segments of protein disorder also exhibited slightly different amino acid preferences[29]. For example, short disordered regions were more depleted in I, V and L, while long disordered regions were more enriched in K, E and P but were less enriched in Q. In addition, long disordered regions were depleted in G and N, while short disordered regions were enriched in G and D [30]. Although these differences were smaller compared to the differences observed between ordered and disordered proteins in general, they highlighted significant heterogeneity within the class of disordered proteins.

Besides amino acid compositional bias, another indication of the unusual sequence properties of disordered proteins is the presence of low complexity regions. These regions often stand out even by simple visual inspection of the sequence, as they usually appear as long stretches containing only one or a few amino acids. This is an indication of low compositional complexity and can be characterized using the concept of sequence entropy. Compositional complexity measures were introduced first for the purpose of sequence alignments and searches and can be viewed as the earliest attempt to identify non-globular proteins[31,32]. Globular proteins have high compositional complexity, very similar to random sequences. In contrast, certain disordered proteins often contain low complexity segments[25], and the more biased the amino acid composition of disordered segments, the more likely it is to be also of low complexity[33]. Nevertheless, the overlap between disordered and low complexity regions is far from complete: 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.

2.2 Databases

For a more detailed understanding of protein disorder, comprehensive databases are needed. This motivated the establishment of the DisProt database[34] that aims to collect disordered proteins and protein regions characterized by various experimental techniques. Entries in this database are collected from the literature and contain at least one experimentally verified disordered region. Detection methods include X-ray crystallography, NMR spectroscopy, CD spectroscopy (both far and near UV) and protease sensitivity, in addition to several other less frequently used experimental techniques. Besides the information about the location of disordered regions, the database contains functional annotations and crosslinks to other databases. As of Release 6.0 (July 1, 2012), DisProt contained 667 proteins containing 1,467 disordered regions.

Another source of disordered proteins is the depository for high resolution structures, the PDB database[35,21]. Although this database is expected to be dominated by ordered proteins, indirectly, it also contains information about protein disorder. In protein structures solved by X-ray crystallography, disorder is defined by missing electron density. In NMR structures, high conformational variability across different NMR models is considered as an indication of disorder. In both cases, disordered residues usually appear within the context of ordered structures, either as terminal regions or short loops within an otherwise ordered protein. The length of these disordered regions spans from a single residue to hundreds, but most often are less than 30 residues long, in contrast to disordered regions in the DisProt database, which are generally longer. Various comparisons indicate that the two databases differ not only in the length of these segments, but encompass two different flavors of protein disorder.

The various datasets are essential components of disorder prediction methods for both optimization and evaluation. During the development of methods, various sequence properties of a compiled dataset of disordered proteins is contrasted to a dataset of globular proteins. It is worth noting that existing datasets of experimentally verified ordered and disordered regions can contain many mis-classified segments. The source of misclassification can be crystals contacts, complex formations or binding of cofactors, all of which can force regions that are flexible in isolation to become structured. Many disordered regions are characterized by semi-quantitative experiments only, lacking position specific information, therefore they are even more prone to misclassification. Furthermore, the order/disorder status can also be sensitive to various environmental conditions[36,37]. The number of known disordered segments is still relatively low and sequence databases are likely to contain many more disordered proteins that are yet uncharacterized. 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.

3 Overview of Protein Disorder Prediction Techniques

The compositional bias of disordered proteins suggests that protein disorder is encoded in the amino acid sequence similarly to the way the folded structure of globular proteins is encoded. This enables the prediction of protein disorder from the amino acid sequence. Currently, more than 50 prediction methods have been published. Some methods utilize machine learning approaches while others are based on simple biophysical considerations. The simplest methods, however, rely on a single amino acid scale[38,19,39]. In general, properties strongly correlating with hydrophobicity, such as flexibility and coordination number, had the highest discriminatory power among various amino acid properties[40,41]. Another property, the tendency of each amino acid to participate in regular secondary structure elements as opposed to be in coil structures, indirectly also correlates with hydrophobicity and is utilized in the Globplot method[42]. The increase in the size of datasets allowed the application of a brute-force approach to directly optimize a specific amino acid scale to discriminate between the two classes. Although in some cases a single effect captured by the amino acid scale is sufficient to explain disorder, generally more sophisticated methods are needed to account for this complex phenomenon.

The field of predicting protein disorder has benefited from the experience of earlier prediction methods developed for various problems in structural biology. In the algorithmic sense, the prediction of protein disorder can be viewed as a classic binary classification problem. Several standard machine learning techniques have been developed and applied for similar problems, such as the prediction of secondary structure, solvent accessibility, functional sites or transmembrane helices. The most commonly used techniques are support vector machines (SVMs) and neural networks. The advantage of machine learning approaches is that they can automatically distill some basic relationships between the input sequence features and the output property. In the specific case of the prediction of protein disorder, the novelty of most methods based on machine learning approaches lies in the representation of input information, rather than in the algorithms themselves. As an input, usually the amino acid sequence within a local sequence window is used. In some cases the amino acid composition or an amino acid propensity within a given window is calculated instead, to reduce the dimensionality of the input data. Some methods also incorporate information about low complexity segments as it can be an important component of a certain type of disorder[33,25].

Additional predicted properties, including secondary structure or solvent accessibility can be also plugged into machine learning techniques[43,44]. However, the benefit from these predictions seems to be much smaller than in some other areas of structure prediction. The likely reason for this is that these methods have been exclusively trained on ordered proteins, and should be used only with caution for disordered proteins. For example, predicted secondary structure does not necessarily contradict protein disorder. Often these regions correspond to transient secondary structural elements, or — in the case of disordered binding regions — to the conformation adopted in the complex form[45]. In the isolated form, with the exception of highly specific scenarios[46], predicted secondary structures are not expected to be stable for disordered proteins.

The incorporation of sequence profiles calculated from evolutionarily related sequences is also more problematic in the case of disordered proteins. The strong sequence bias present in these proteins, especially in low complexity segments can distort the result of sequence similarity searches. Generally, disordered proteins are evolutionarily less conserved[47], but the dynamic behavior and the associated molecular function can be preserved even in the absence of apparent sequence conservation[48]. As a result, alignments are a less reliable source of information for disordered protein segments. Although several methods use evolutionary information in the prediction, it leads to a smaller boost in the performance of disorder prediction methods than observed for example in the case of secondary structure prediction methods[49].

Most prediction methods provide predictions at the per residue basis. The performance of disorder predictions can be evaluated using the Matthews correlation coefficient (MCC), balanced accuracy (Acc) that weights the performance on the positive and negative datasets based on the respective size of the datasets, and the area under the receiver operating characteristic (ROC) curve (AUC, with possible values ranging from 0.5 for random predictions to 1.0 for perfect predictors).

Since 2002, the performance of various disorder prediction methods has been critically assessed at the CASP experiments[50-54]. According to the latest evaluation[54], top methods can reach 0.85 AUC. In other terms, they can identify around 75% of disordered residues at the expense of misclassifying around 25% 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. Testing on disordered regions culled from DisProt usually place different methods at the top. On these datasets, methods can discriminate between ordered and disordered segments with around 80% accuracy at the per residue basis[55,49,56,57]. However, several methods have been trained on this dataset, therefore these numbers should be treated with caution. Generally, the performance of disorder predictors critically depends on the dataset used for testing, or more generally, the type of disorder studied. It is also influenced by the evaluation criteria. Nevertheless, modern disordered prediction methods can be considered quite reliable in general.

3.1 Machine Learning Methods

A comprehensive review of published methods appeared in the literature recently[58]. The exhaustive enumeration of all present algorithms is beyond the scope of this chapter, instead our aim is to cover the basic approaches in this field. 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 at the end of the section.

The first method developed for the prediction of disordered proteins is PONDR VL-XT[33]. The training set of this method was composed of variously characterized long (> 30 residues) disordered regions[59], and two additional training sets of X-ray-characterized terminal regions, one for the amino-terminus and one for the carboxy-terminus[28]. The method uses the amino acid compositions, attributes derived from compositions such as sequence complexity, and attributes derived from compositions via some function or scale such as hydropathy, net charge, etc. The attributes were selected by analyzing their discriminatory power, their orthogonality and based on their effect on the performance. Then, the various types of attributes were weighted and combined via artificial neural networks (ANNs). The resulting method was found particularly useful to pinpoint certain regions that are candidates for undergoing disorder-to-order transitions[60,61].

Another member of the PONDR family of predictors is VL3[62]. It also uses an artificial neural network but the training dataset was much larger compared to that of VL-XT. The input is formed by 18 amino acid frequencies, the average flexibility and sequence complexity, calculated within a window of 41 residues. Similarly to VL-XT, a neural network with a fully connected hidden layer of ten neurons was trained on the specific datasets and it outputs a value for the central amino acid in the window. Homologous sequences were also included in the training set to increase the number of examples. Sequence profiles generated by PSI-BLAST can also be added as an input attribute to improve the accuracy of predicting disordered regions.

DisEMBL, another computational tool for predicting disordered/unstructured regions was developed by Linding et al[63]. Because of some uncertainties in the definition of protein disorder, they developed three separate neural network based predictors using alternative definitions of disorder. These correspond to missing residues indicated by REMARK 465 in the PDB files, residues with high B-factor (hot loops) and residues within loops and coils. The differences in these three predictors underlined the distinct features of each group. By investigating the relationships between the different disorder definitions, they found that hot loops showed less correlation with coils and more with the missing residues.

DisPSSMP[64] uses a radial basis function network as a training algorithm. The input of the method is calculated from position specific scoring matrices generated by PSI-BLAST[65] that are 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 also incorporates secondary structure predictions and introduces a two-stage classifier to further enhance the prediction power. In the second stage, the predictions are smoothed and refined by adjusting the threshold value and the size of a sliding window based on the outputs of the first layer.

Using an original approach, RONN[44] recognizes disordered segments based on their similarity to well-characterized prototype sequences with known disordered status. 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 converted into distances and are used to train a modified version of radial basis function networks called a bio-basis function neural network.

Along with artificial neural networks, the most widely used class of standard machine learning algorithms are support vector machines (SVMs). SVMs have several advantages over neural networks as they are less prone to overfitting, can be trained more efficiently and handle noisy datasets better. SVMs can also handle unbalanced datasets, which is the case for disordered residues defined based on missing residues, as these usually comprise only 10% of all residues. The first method utilizing SVMs for the prediction of disorder was implemented in DISOPRED2[7]. This method was trained on a large dataset of missing residues of high resolution structures. Separate models were created for N- and C-terminal regions besides the model for the middle regions of the sequences. The input of the predictions is a sequence profile for each protein, generated using a PSI-BLAST search[65] against a filtered sequence database. One of the keys of the high accuracy of DISOPRED2 was that it was trained by placing a larger cost on false positive predictions.

PrDOS is a hybrid method that combines an SVM-based prediction method with a template based prediction[66]. For each query sequence, a position-specific scoring matrix is generated after two-rounds of PSI-BLAST searches against a non-redundant sequence database. The profiles are used for a template based search to find a potential homologe with known status of order or disorder in the PDB. If no such case is found, an SVM-based prediction is carried out based on local sequence windows of the profiles.

In the case of feed forward neural networks and SVMs, the prediction for each residue is independent of the prediction for other residues. In contrast, recurrent networks can also propagate data from later processing stages to earlier stages. Such technique is used in DISpro[43]. It employs a one-dimensional recursive neural network that combines the flexibility of a Bayesian model with the fast and convenient parameterization of neural networks. The method also incorporates 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.

Another approach that can take into account the predicted disorder tendency of neighboring positions was recently published. The OnD-CRF method[67] utilizes conditional random fields for the prediction of protein disorder. The method relies on features generated from the amino acid sequence and from secondary structure prediction. The training data set was derived from high-resolution crystal structures that lack coordinates for those amino acids that are considered to be disordered, and the performance was optimized with respect to the area under the ROC curve.

Instead of using explicit datasets for disordered proteins, methods can also exploit the information stored in large sequence databases. The DRIP-PRED method[68] 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. 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.

The methods described so far are all specific to one type of protein disorder only, represented either by the DisProt[34] dataset or missing residues of X-ray structures. Their performance tested on the other dataset resulted in significantly lower efficiencies. This problem was addressed by the PONDR VSL2 method[49,69]. 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. The metapredictor 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.

POODLE-I[70] also integrates methods that target different disordered regions according to their length, by incorporating specific predictors that recognize short and long disordered segments as well as mostly disordered proteins. It also assumes that the factor causing a short disordered region might be different from the factor causing a long one: a short disordered region, such as a loop or linker is mainly determined according to whether it is located within an otherwise well defined structure. By contrast, the formation of a long disordered region is mainly affected by the physico-chemical properties derived from the sequence such as low hydrophobicity or high charge content. Accordingly, POODLE-I is based on a workflow approach for combining prediction results from the POODLE series, the elements of which are individual prediction algorithms. Additionally, POODLE-I uses predicted structural information as well, such as secondary structure.

Meta approaches that integrate the results of several prediction methods have been very successful in various areas of structure predictions[71] and appeared for the prediction of protein disorder as well. These methods achieve improved performance by decreasing the noise of individual predictors. Since individual disorder prediction methods are often specific to certain types of protein disorder, their combination could cover more aspects of disorder. The last rounds of CASP experiment were clearly dominated by meta-predictors[54]. 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 insights into the structural and even the functional properties of disordered regions.

3.2 Incorporating Physical Principles into Disorder Prediction

As opposed to the application of various ‘black box-like’ machine learning algorithms, the prediction of protein disorder can be approached with 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[38,19]. 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[74] to provide a position specific prediction. A similar concept is behind the FoldUnfold method[39]. It predicts proteins 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.

Taking one step further, modeling of residue-residue interactions can be incorporated into the prediction of protein disorder. A prime example of the more sophisticated physics-based methods is the IUPred algorithm[73]. This method captures the essential cause of protein non-folding: if a residue in a protein is not

Table 1 Summary of the 13 analyzed disorder prediction methods. Column 2 shows the dataset on which the methods were trained, column 3 shows the basic implemented algorithm and column 4 shows the quantities the algorithm uses to calculate the final prediction score. Abbreviations: SVM – support vector machine; SOM – Self organizing map; SGT – spectral graph transducer; PSSM – position specific scoring matrix.

Name of method	Training dataset	Algorithm	Input data of the algorithm
VL-XT[33]	XT: missing residues in X-ray structures (terminal regions) VL:variously characterized long disordered segments	neural network	amino acid frequencies, amino acid propensities
VL3[62]	variously characterized long disordered segments	neural network	amino acid frequencies, amino acid propensities, sequence complexity
DisEMBL[63]	missing residues in X-ray structures	neural network	single sequence window
DisPSSMP[64]	DisProt	radial basis function neural network	PSI-BLAST PSSM condensed by physico-chemical properties, secondary structure prediction
RONN[72]	missing residues in X-ray structures	bio-basis function neural network	single sequence window
DISOPRED2[7]	missing residues in X-ray structures	SVM and neural network	PSI-BLAST PSSM windows
PrDOS[66]	missing residues in X-ray structures	SVM and template based prediction	PSI-BLAST and homologous structures
DISpro[43]	missing residues in X-ray structures	1D recursive neural network	full length PSI-BLAST PSSM, secondary structure and solvent accessibility prediction
OnD-CRF[67]	missing residues in X-ray structures	conditional random fields	single sequence, secondary structure prediction
DRIP-PRED[68]	UniProt sequences	Kohonen SOM	PSI-BLAST PSSM windows, secondary structure prediction
VSL2B[69,49]	missing residues in X-ray structures and DisProt	SVM	amino acid propensities, sequence complexity, (PSI-BLAST PSSM) (secondary structure prediction)
POODLE-I[70]	missing residues in X-ray structures, DisProt and SwissProt	SVM and SGT	amino acid propensities, PSI-BLAST PSSM, secondary structure prediction
IUPred[73]	none	biophysical model	amino acid composition

able to form enough favorable intrachain contacts, it will not adopt a stable position in the 3D structure of the chain. 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 in IUPred is done taking an energetics point of view. For globular proteins, the contribution of interresidue interactions to total energy is often approximated by low-resolution force fields, or statistical potentials, which are energy-like quantities derived from globular proteins based on the observed amino acid pairing frequencies[75]. In deriving the actual potentials, different principles have been applied. The resulting empirical energy functions are well suited to assess the quality of structural models and have been used for fold recognition or threading but also in docking, ab initio folding, or predicting protein stability. Their success in a wide range of applications suggests the existence of a common set of interactions, simultaneously favored in all native — as opposed to alternate — structures.

In the case of IUPred, a dedicated statistical potential is optimized to estimate the pairwise interaction energies between residues. The total pairwise energy *E* of a protein in its native state is the sum of the energies of all the pairwise residue-residue interactions in the protein. *E* is the function of the conformation as well as the amino acid sequence, as they define the list of residue-residue interactions that have a contribution to the total energy. This total energy can be calculated by taking all contacts in the protein, and weighting them by the corresponding interaction energies. The interaction energy between any two types or amino acids can be inferred by calculating the frequency of interactions between these two types in a dataset of known protein structures. These frequencies are transformed into interaction energies using the Boltzmann hypothesis[76] and are described by the 20 by 20 interaction energy matrix of amino acid pairs, **M**. Hence, the pairwise energy content calculated based on the structure can be written as:

$$E_{calculated} = \sum_{i,j} M_{ij} C_{ij} \tag{1}$$

where M_{ij} is the interaction energy between amino acid types i and j, and C_{ij} is the number of interactions between residues of types i and j in the given conformation.

This energy calculation, however, assumes the knowledge of the 3D structure of the protein and as such, is not directly applicable to proteins whose structure can not be determined. To come around this problem, a novel estimation scheme was established and implemented in IUPred to enable the estimation of the *E* interaction energy without the structure, using the protein sequence alone. The rationale behind this approach is that the energy contribution of a residue depends not only on its amino acid type, but also on its potential partners in the sequence. It is assumed that if the sequence contains more amino acid residues that can form favorable contacts with the given residue, its expected energy contribution will be more favorable. The simplest approximating formula for the specific estimated pairwise energy can be expressed with a quadratic formula as:

$$E_{estimated} = L\sum_{i,j} P_{ij} f_i f_j \qquad (2)$$

where L is the length of the protein, f_i is the normalized frequency of residues of type i and **P** is the energy estimator matrix. The elements of **P** are optimized on a set of globular proteins using the least squares method in order to minimize the difference between $E_{calculated}$ and $E_{estimated}$. Equation (2) gives an estimate for the energy of the whole protein, however, it can be naturally modified to calculate the pairwise energy of single residues as well. For this, it has to be considered that in multi-domain proteins the residues belonging to different domains do not interact. For this reason, for each residue the amino acid frequencies are only calculated in the sequential neighborhood roughly corresponding to the average domain size. The width of this sequence window is marked by w_0 and is set to 100 residues to each side. To estimate the interaction energy of residue k (of type j), equation (2) can be modified:

$$E_j^k = \sum_{i=1}^{20} P_{ij} f_i^k (w_0) \qquad (3)$$

where $f_i^k(w_0)$ is the fraction of residues of type i in the w_0 neighborhood of residue k. (Note that lower indices stand for amino acid type, while upper indices stand for position in the chain.) Formula (3) enables the estimation of the intrachain interaction energies of each residue directly from the amino acid sequence. Generally, residues with less favorable predicted energies are more likely to be disordered. Testing on 559 globular and 129 disordered proteins[73] showed that this energy estimation scheme is accurate enough to achieve a high true positive rate (fraction of disordered residues correctly predicted) of 76% while maintaining a sufficiently low false positive rate (fraction of ordered residues incorrectly predicted) of 5% — a standard choice of type I error in prediction methods. The strength of the construction of the method is that its parameters 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.

The above energy estimation method is implemented in IUPred. The method is accessible via a web server[77] hosted at the Institute of Enzymology (http://iupred.enzim.hu). For the ease of interpretation, the calculated energies are converted into probability values, indicating the probability of each residue being disordered. Figure 2 shows an example output of the IUPred server for the human Wiskott-Aldrich protein (WASp). WASp is a 502 residue long protein that is entirely disordered with the exception of the ordered WH1 domain spanning the

39-148 region. The assigned probabilities are in accordance with the known structural information as the calculated probabilities on the ordered domain lie below 0.5 marking order (low probability of being disordered) and above 0.5 for the rest of the protein (high probability of being disordered).

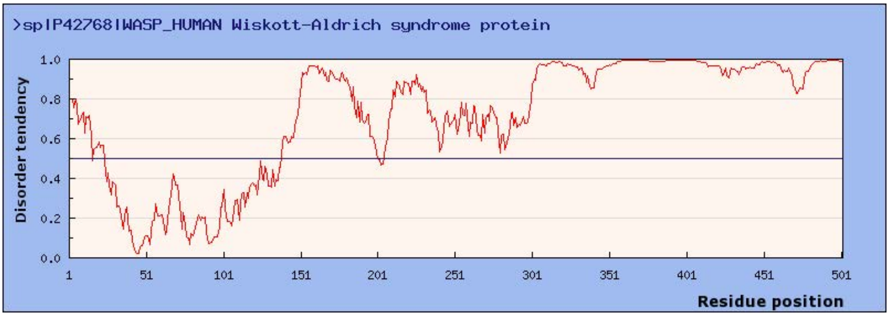


Fig. 2 Screenshot of the IUPred server output for the human Wiskott-Aldrich protein. The horizontal axis represents the protein chain and the vertical axis represents the probability of each residue to be disordered. Residues with values above 0.5 are predicted to be disordered and values below 0.5 indicate an ordered structure.

4 Prediction of Disordered Binding Regions

As discussed in section 1.2, 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 proteins and disordered proteins in general, and these properties — in principle — enable the construction of prediction algorithms to recognize them from the protein sequence. 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[6] aims to capture the basic biophysical properties of disordered binding segments. 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. In this model the combination of the high disordered tendency of the sequential 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 these criteria for the presence of a disordered binding region are quantified with the use of estimated energies.

The testing of ANCHOR showed that the predictor recognizes 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, acidic 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 a predictor. This independency also shows the generality of ANCHOR. The method combines the transparency of simplified biophysical models with the usability of bioinformatical approaches.

The predictions obtained with IUPred and ANCHOR are demonstrated through the example of the human calcium/calmodulin-dependent protein kinase IV (UniProt ID: Q16566), shown on figure 3. The plot was generated with the online version of ANCHOR[78], available at http://anchor.enzim.hu/. Calcium/calmodulin-dependent kinase IV binds to calmodulin near its C-terminal end (residues 322-341). 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 consisting of three positively charged residues followed by three hydrophobic ones in the 1st, 8th and 14th position C-terminal from the positive sequence patch. The location of this motif is also indicated on the figure.

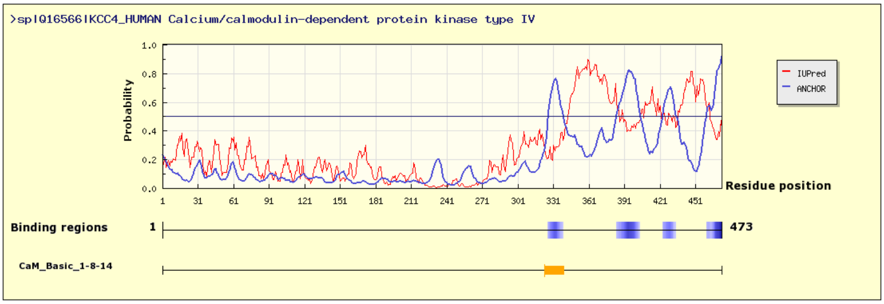


Fig. 3 Output of the ANCHOR prediction server for calcium/calmodulin-dependent protein kinase IV. The plot shows the predicted disordered binding regions in blue with the output of the general prediction method IUPred in red and the location of the calmodulin binding motif in orange.

Although IUPred and ANCHOR rely on the same approach and use the same interaction energy prediction scheme, their outputs are distinctively different. However, IUPred also reacts to the presence of disordered binding regions: as can be seen from the example presented on figure 3, disordered binding regions tend to appear more ordered than their surrounding disordered protein segments. This tendency is not exclusive to IUPred, many other disorder prediction outputs reflect binding regions in a similar way. In the case of PONDR VL-XT, the presence of these 'dips' in the prediction profile was exploited to construct a disordered binding region prediction algorithm[60,61]. In this framework, regions undergoing a coupled folding and binding process adopting an α-helical conformation in their bound form were targeted. These regions, termed α-MoRFs were predicted using the local drops in the prediction score as an input to a neural network that was trained on known examples of α-helical binding sites. The neural network then tries to discriminate the potential binding regions using various sequence features, including disorder, secondary structure predictions and amino acid indices.

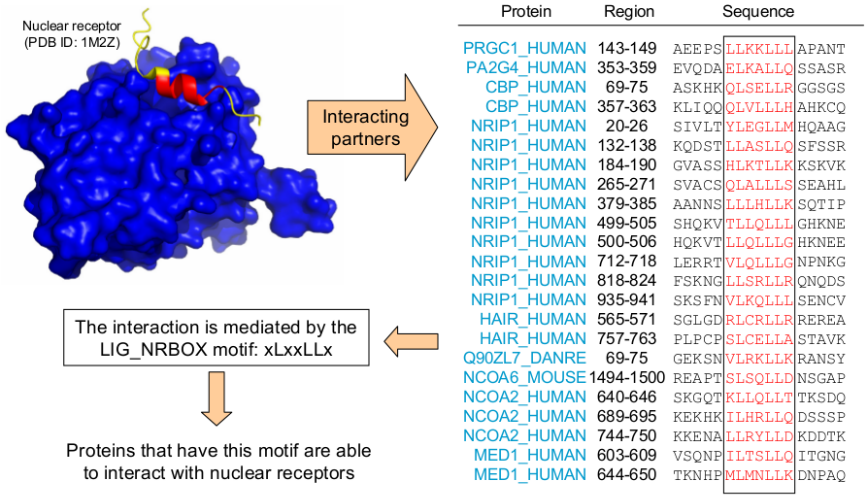
5 Linear Motifs

As discussed in the previous section and section 1.2, the study of protein-protein interactions formed by disordered proteins is based on structural considerations. However, the study of interactions between protein domains and short, linear protein regions — a description which fits most interactions between folded and disordered proteins — has a distinctively separate approach as well, with the use of linear motifs.

5.1 Defining and Using Linear Motifs

In the framework of linear motifs, the interaction is not described focusing on the short partner, but the large one, which is usually a protein domain. It was found for many domains such as SH2/SH3, 14-3-3, WW and kinase domains that their interacting partners — albeit in many cases not being homologues — share a limited number (typically between 2–10) of common residues in the short interaction region[79,80]. Apart from these residues, the binding region also incorporates other, flexible positions that can contain various amino acids without disrupting the binding[81]. Figure 4 shows the example of nuclear receptors that are able to bind a large variety of protein partners. Although most partner proteins are not homologues, they all share three key leucine residues at their interacting sites. During the interaction, the region that binds to the receptor forms an α-helix and the three leucines form a hydrophobic patch on the surface of the helix. This patch in turn recognizes the appropriate complementary hydrophobic region of the interface of the receptor and anchors the helix to the binding groove. The consensus sequence of the binding region is xLxxLLx, where x can stand for any amino acid, except for proline, as it would disrupt the helix formation. This motif is called

LIG_NRBOX and ligands of many nuclear receptors are able to recognize their partners via this sequence pattern. The theory of linear motifs, used to describe such interactions, is based on the assumption that these common residues (constituting the motif) mediate the binding largely independent of the other regions of the protein they are embedded in, functioning autonomously. However, in many cases the role of the context was shown to be larger than originally expected[82].


Nuclear receptor
(PDB ID: 1M2Z)
Interacting partners
The interaction is mediated by the LIG_NRBOX motif: xLxxLLx
Proteins that have this motif are able to interact with nuclear receptors

Protein	Region	Sequence		
PRGC1_HUMAN	143-149	AEEPS	LLKKLLL	APANT
PA2G4_HUMAN	353-359	EVQDA	ELKALLQ	SSASR
CBP_HUMAN	69-75	ASKHK	QLSELLR	GGSGS
CBP_HUMAN	357-363	KLIQQ	QLVLLLH	AHKCQ
NRIP1_HUMAN	20-26	SIVLT	YLEGLLM	HQAAG
NRIP1_HUMAN	132-138	KQDST	LLASLLQ	SFSSR
NRIP1_HUMAN	184-190	GVASS	HLKTLLK	KSKVK
NRIP1_HUMAN	265-271	SVACS	QLALLLS	SEAHL
NRIP1_HUMAN	379-385	AANNS	LLLHLLK	SQTIP
NRIP1_HUMAN	499-505	SHQKV	TLLQLLL	GHKNE
NRIP1_HUMAN	500-506	HQKVT	LLQLLLG	HKNEE
NRIP1_HUMAN	712-718	LERRT	VLQLLLG	NPNKG
NRIP1_HUMAN	818-824	FSKNG	LLSRLLR	QNQDS
NRIP1_HUMAN	935-941	SKSFN	VLKQLLL	SENCV
HAIR_HUMAN	565-571	SGLGD	RLCRLLR	REREA
HAIR_HUMAN	757-763	PLPCP	SLCELLA	STAVK
Q90ZL7_DANRE	69-75	GEKSN	VLRKLLK	RANSY
NCOA6_MOUSE	1494-1500	REAPT	SLSQLLD	NSGAP
NCOA2_HUMAN	640-646	SKGQT	KLLQLLT	TKSDQ
NCOA2_HUMAN	689-695	KEKHK	ILHRLLQ	DSSSP
NCOA2_HUMAN	744-750	KKENA	LLRYLLD	KDDTK
MED1_HUMAN	603-609	VSQNP	ILTSLLQ	ITGNG
MED1_HUMAN	644-650	TKNHP	MLMNLLK	DNPAQ

Fig. 4 The figure shows the known interaction partners of nuclear receptors that all bind using the same binding mode. The upper left structure shows a solved complex structure (based on PDB entry 1m2z) between a small region of the human NCOA2 nuclear receptor coactivator (shown in red and yellow) and a glucocorticoid receptor (shown in blue). Although the actual sequences around the binding region do not share a high level of similarity, they all contain three key leucine residues. These three amino acids interspersed and flanked by flexible positions constitute the consensus LIG_NRBOX motif (shown in red in the structure and the partner sequences).

The majority of protein-protein interaction mediating linear motifs were described in eukaryotes. Currently, the largest and most comprehensive available database of these motifs is the Eukaryotic Linear Motif (ELM) database[83]. Motifs are categorized according to the type of interaction partners and functions (cleavage sites, generic protein-protein interaction sites, post-translational modification sites and targeting signals). Although the majority of these motifs were described in eukaryotic proteins, some of them can be expected to occur in proteins of bacteria and archaea too. Furthermore, instances of the retinoblastoma protein-, the SH3- and the 14-3-3 interacting motifs, among others, were identified in various viruses as well[84].

Linear motifs not only serve as a simplified description of a protein-protein interaction mode, but serve as a prediction algorithm too. Consensus motifs can be readily used to search for binding partners of a given domain in unknown sequences through basic pattern matches. The strength of this method besides its simplicity is that it automatically gives information about the possible interacting partner. However, these patterns usually consist of only a few fixed residues, and therefore most motifs are weakly defined, meaning that matches can arise purely by chance with a relatively high probability[85]. As a result, naïve motif searches are hindered by the massive amount of false positive hits. This is partially the result of the incomplete description sequence patterns offer. Inside a living cell, the functionality of the motifs is modulated by structural, spatial and temporal control[86]. Furthermore, the proper structural context of a motif (such as being accessible, flexible and capable of forming the secondary structure necessary to fit into the binding cleft of the target domain) is crucial for its biological relevance and motif definitions do not include any such information.

5.2 Linear Motifs and Disordered Binding Regions

The disordered binding region and the linear motif concepts describe molecular interactions on different bases: the former focusing on the structure (or the lack and formation of it) and the latter approaching the problem through the sequence. However, the interactions described by the two concepts share a high degree of similarity. In both cases the interaction is confined to a relatively short, linear sequence region in one of the partners. Furthermore, most experimentally described linear motif instances were found in disordered protein regions. Accordingly, in many cases, such as the binding of p53 to MDM2 and the N terminal region of p27 binding to the cyclinA-CDK2 complex, the same interaction was categorized as an example of both linear motif mediated binding and of disordered binding regions. Through many common examples, both the binding of disordered proteins and linear motifs have been shown to play vital roles in eukaryotic regulation and signaling[86]. This, however, also serves as a potential point of attack for many successful viruses (such as HIV or ebola) that also harbor disordered proteins containing various motifs[84]. Apart from individual examples, the connection between protein disorder and motif regulation has been also shown at a more general level[87].

Despite the very different approaches used to describe interactions via disordered binding regions and linear motifs, the two fields not only share a large number of common examples but also struggle with essentially the same problems. Probably the most serious bottleneck in both cases is the low number of experimentally verified examples. About 50% of human proteins are predicted to contain at least one larger disordered region, and it was shown that the primary reason for the emergence of these regions is to harbor binding regions[6]. In contrast, the number of experimentally verified disordered regions collected in the DisProt database is in the hundreds[34] and the number of known disordered binding regions is even lower. Parallelly, recent results providing a moderate estimate places the number of individual motif mediated interactions in the human proteome above 35,000[88]. Despite this high estimated occurrence, the number of experimentally verified, true motif instances in all eukaryotic proteins described in the ELM database has yet to reach 2,000. While it is clear that the two concepts — linear motifs and disordered binding regions — could be used in connection to strengthen each other's predictions, this connection between the two fields is yet to be established in detail.

6 Using Predictions on Disordered Proteins – A Practical Guide

6.1 How to Use Disorder Prediction Methods

Disorder prediction methods can be used in two different ways. On one hand, they can be used in large-scale studies where many proteins are analyzed. These projects usually aim to uncover statistically meaningful differences between classes of proteins of the proteomes of different organisms with regard to disorder content. In this scenario usually only longer, contiguous disordered segments are considered, and short runs (typically below 20 or 30) of residues predicted to be disordered are filtered out. In this setup, methods that are trained to recognize longer stretches of disordered residues, such as IUPred, RONN, DisPSSMP or PONDR VL3 clearly have an advantage. Although practically all state-of-the-art methods assign a continuous score to each residue, representing the probability of it being disordered, when using these methods, this score is converted to a binary classification. Residues with scores above a pre-determined threshold are classified as disordered, and residues with lower scores are assigned an ordered status. It is worth noting, however, that various methods are optimized for different false positive prediction rates — usually in the 2–15% range — and the pre determined cutoff is set accordingly. Although in comparative studies, where the basic questions are similar to "which of these groups of proteins contains more disorder" or "how does the disorder content of proteomes change during evolution" this does not affect the final results to a great extent, it should be kept in mind that the actual numbers depend on the choice of algorithm.

The other typical use of disorder prediction methods is the analysis of individual proteins. In these cases, the difference between the false positive rates of various methods presents a clear disadvantage, as the choice of method clearly affects the results. Although this in theory can be circumvented by recalibrating various methods on a standardized dataset, this solution is not feasible for casual users. Furthermore, the fact that various methods are optimized for various typical lengths of disorder presents an additional level of difficulty when choosing a single method to use. These considerations point toward the combined use of disorder prediction methods when investigating individual protein sequences. A good starting point can be the application of methods sensitive to larger, contiguous regions of order/disorder to establish the basic structural composition of the protein in question. As a next step, methods capable of detecting more localized disorder regions — such as DISpro or DISOPRED2 — can be applied.

Probably one of the most difficult tasks from the viewpoint of successful disorder prediction is presented by partial or transient structural elements. In the case of stable, globular domains, or highly flexible disordered regions without a strong structural preference, most methods tend to show good agreement. However, considering regions with partial or transient structure, such as molten globules, coiled-coil regions or some disordered binding regions, almost all methods react to the underlying structural preferences with a lowered prediction score[58]. This type of behavior and the resulting lack of a clear consensus prediction is highly characteristic of these structurally ambiguous regions and for the experienced researcher these can serve as dead giveaways. However, in the successful identification of the nature of the underlying structural reasons, dedicated predictions — such as ANCHOR for identifying disordered binding regions or COILS[89] for the identification of coiled-coil regions — are indispensable.

In the next section we present a case study, where the reaction of various prediction methods are demonstrated for ordered, disordered and disordered binding regions of the human p53 protein.

6.2 Bringing It All Together – An Application to p53

In this section, we show an application of the principles described in the previous section through the example of human p53. p53 is a 393 residue long tumor suppressor protein involved in the control of cell-cycle and apoptosis. The protein has a relatively complex architecture containing a central, ordered DNA binding domain (DBD) and two long disordered regions on both sides of the DBD, harboring several binding regions and a tetramerization region. As both the binding regions and the tetramerization region are disordered in isolation but can adopt a structure upon binding, there is no single good answer for these regions from the perspective of disorder predictions.

Figure 5 shows the output of the different applied prediction methods on the full length of p53. In the central, ordered domain (spanning residues 102–292, marked with red box) virtually all methods agree, assigning a relatively low score to the majority of the domain, indicating the presence of a long region with high structural content. This prediction is in accordance with the results obtained from a secondary structure prediction, indicating numerous β-strands in the domain region. The validity of the predicted ordered region and the type of assigned secondary structures can be ascertained through the solved structure of the DBD. The predicted secondary structures correspond to the experimentally measured structure with a relatively high precision.

The outputs of various methods on the N-terminal disordered region (encompassing residues 1–101) are much more heterogeneous in comparison. This regions is known to contain (at least) 3 different disordered binding regions: the segment between residues 17–27 binds to MDM2, the other two binding sites overlap with residues 33–56 binding to RPA 70N and residues 45–58 binding to the B subunit of RNA polymerase II (all shown with green boxes). Basically all methods react to the presence of the binding regions with a lower score, however, to a varying degree. Some methods, such as VSL2B, VL3 and POODLE-I predict the whole region to be disordered. On the other extreme, DISpro predicts this region to be completely ordered with a very low score. However, some methods, such as IUPred, RONN, DisPSSMP and OnD-CRF react to the presence of transient structure by assigning a score very close to 0.5 which effectively corresponds to a 'non-prediction': these methods realize that in the binary framework of 'ordered or disordered' they cannot correctly classify these regions. Some methods, such as VSL2B, POODLE-I or DISpro give a general indication of the underlying structural tendency by giving one extended dip covering the whole interacting region. Others, such as DRIP-PRED, DISOPRED2, IUPred or PrDOS give two distinct dips corresponding to the MDM2 binding region and the other two, overlapping regions. This behavior is also characteristic of VL-XT being highly sensitive to local structure, which shows in that this is the only method that scores the MDM2 binding region with a significantly lower score than the other binding region. It is worth noting that the MDM2 binding site has a slight α-helical tendency even in the unbound form and this helix is stabilized via the interaction. This structural tendency is also shown by the secondary structure prediction by PSI-PRED[90]. Furthermore, the MDM2 binding region also contains the MDM2 interaction linear motif, giving further support to the predictions and hinting at the interaction partner. However, the strongest prediction-level evidence hinting at the presence of binding regions (as opposed to coiled-coil region or a short collapsed structure) is the high-confidence predictions of ANCHOR covering all three binding regions.

The C-terminal disordered region (from 293–393) is structurally reminiscent of the N-terminal region. It is generally disordered and contains multiple, overlapping binding regions. As in the case of the N-terminal region, there is a high consensus between different prediction methods concerning the non-interacting disordered regions. In the tetramerization region (residues 325-356) all methods exhibit a lower score but again — similarly to the N-terminal binding regions — to a highly varying degree. The assigned scores range from the clearly disordered prediction of VSL2B to the low scores of DRIP-PRED and DISpro predicting the region to be ordered. However, IUPred, RONN, DisPSSMP and OnD-CRF again give a score close to 0.5 indicating their justified inability to give a definite prediction. The presence of a binding region is again supported by the high confidence ANCHOR prediction and the PSI-PRED prediction gives an indication of the α-helical structure adopted in the bound form. Apart from the tetramerization site, the C-terminal region also contains a binding region that is able to bind to a multitude of different partners acting as a molecular switch. The prediction algorithms consistently react to this region in a fashion similar to the previous binding sites.

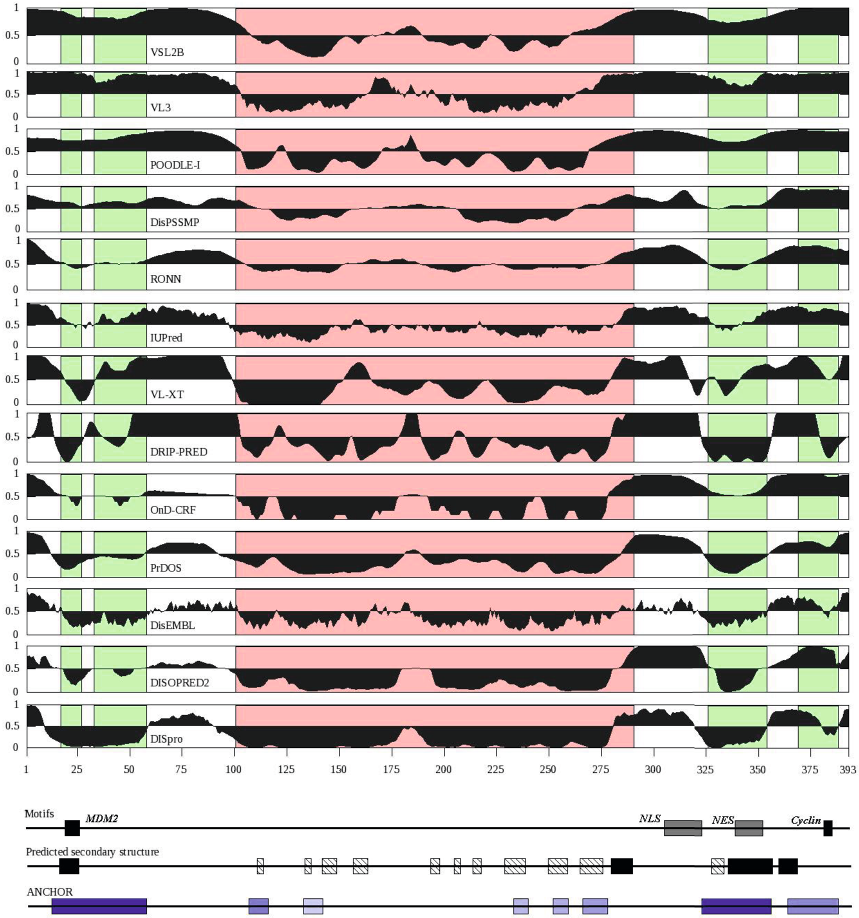


Fig. 5 Predictions for human p53 (UniProt AC: P04637). In the case of DisPSSMP, OnD-CRF and DISOPRED2 the original prediction scores were rescaled linearly to be directly comparable with other methods. Disordered predictions were sorted top to bottom by decreasing average predicted disorder tendency. The central, ordered DNA binding domain is shown in red and experimentally verified disordered binding regions are shown in green while the rest of the protein is disordered and is shown in white. Underneath the disorder prediction outputs, the known biologically relevant linear motifs are shown with black and grey boxes for ligand binding and sub-cellular localization target motifs, respectively. The middle line (Predicted secondary structure) shows the secondary structure prediction by PSI-PRED, black and striped boxes indicating predicted α-helical and β structures, respectively. 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 darker shades indicating a higher confidence.

The more pronounced α-helical preference of the cyclin binding site (embedded in this binding region) can also be seen in many prediction outputs and the secondary structure prediction; the presence of the cyclin binding linear motif and the positive ANCHOR prediction all provide further support to the presence of this interaction. However, the rest of this combined binding region lacks any predicted secondary structure which faithfully reflects the fact that this region is able to bind to its various partner proteins in all three basic secondary structure (α, β and irregular).

The example of p53 shows that the outputs of individual disorder prediction methods can be misleading or difficult to interpret on their own. However, the combination of various methods coupled with other types of structural/functional predictions — such as secondary structure prediction, linear motif searches or disordered binding site prediction by ANCHOR —, can give a detailed and reliable profile for proteins with even highly complex structural features. This example faithfully reflects that upon studying a single protein, the combination and proper interpretation of various predictors can go a long way.

Acknowledgments. This work was supported by the grant Hungarian Science Research Fund (OTKA) NK 100482. The János Bolyai Research Scholarship of the Hungarian Academy of Sciences for Z.D. and C.M. are also gratefully acknowledged. The authors would like to thank Mark Adamsbaum for his critical comments on the text.

References

1. Wright, P.E., Dyson, H.J.: Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. J. Mol. Biol. 293(2), 321–331 (1999), doi:10.1006/jmbi.1999.3110, S0022-2836(99)93110-8 [pii]
2. Dunker, A.K., Lawson, J.D., Brown, C.J., Williams, R.M., Romero, P., Oh, J.S., Oldfield, C.J., Campen, A.M., Ratliff, C.M., Hipps, K.W., Ausio, J., Nissen, M.S., Reeves, R., Kang, C., Kissinger, C.R., Bailey, R.W., Griswold, M.D., Chiu, W., Garner, E.C., Obradovic, Z.: Intrinsically disordered protein. J. Mol. Graph. Model. 19(1), 26–59 (2001), doi:S1093-3263(00)00138-8 [pii]
3. Dyson, H.J., Wright, P.E.: Intrinsically unstructured proteins and their functions. Nat. Rev. Mol. Cell. Biol. 6(3), 197–208 (2005), doi:nrm1589, [pii] 10.1038/nrm1589
4. Tompa, P.: Intrinsically unstructured proteins. Trends Biochem. Sci. 27(10), 527–533 (2002), doi:S0968-0004(02)02169-2 [pii]
5. Dunker, A.K., Obradovic, Z., Romero, P., Garner, E.C., Brown, C.J.: Intrinsic protein disorder in complete genomes. In: Genome Inform. Ser. Workshop Genome Inform., vol. 11, pp. 161–171 (2000)
6. Meszaros, B., Simon, I., Dosztanyi, Z.: Prediction of protein binding regions in disordered proteins. PLoS Comput. Biol. 5(5), e1000376 (2009), doi:10.1371/journal.pcbi.1000376
7. Ward, J.J., Sodhi, J.S., McGuffin, L.J., Buxton, B.F., Jones, D.T.: Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J. Mol. Biol. 337(3), 635–645 (2004), doi:10.1016/j.jmb.2004.02.002, S0022283604001482 [pii]

8. Xie, H., Vucetic, S., Iakoucheva, L.M., Oldfield, C.J., Dunker, A.K., Uversky, V.N., Obradovic, Z.: Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions. J. Proteome Res. 6(5), 1882–1898 (2007), doi:10.1021/pr060392u
9. Tompa, P.: The interplay between structure and function in intrinsically unstructured proteins. FEBS Lett. 579(15), 3346–3354 (2005), doi:S0014-5793(05)00424-2, [pii] 10.1016/j.febslet.2005.03.072
10. Galea, C.A., Wang, Y., Sivakolundu, S.G., Kriwacki, R.W.: Regulation of cell division by intrinsically unstructured proteins: intrinsic flexibility, modularity, and signaling conduits. Biochemistry 47(29), 7598–7609 (2008), doi:10.1021/bi8006803
11. Uversky, V.N., Oldfield, C.J., Dunker, A.K.: Intrinsically disordered proteins in human diseases: introducing the D2 concept. Annu. Rev. Biophys. 37, 215–246 (2008), doi:10.1146/annurev.biophys.37.032807.125924
12. Cheng, Y., LeGall, T., Oldfield, C.J., Dunker, A.K., Uversky, V.N.: Abundance of intrinsic disorder in protein associated with cardiovascular disease. Biochemistry 45(35), 10448–10460 (2006), doi:10.1021/bi060981d
13. Uversky, V.N.: Intrinsic disorder in proteins associated with neurodegenerative diseases. Front Biosci. 14, 5188–5238 (2009), doi:3594 [pii]
14. Uversky, V.N., Oldfield, C.J., Midic, U., Xie, H., Xue, B., Vucetic, S., Iakoucheva, L.M., Obradovic, Z., Dunker, A.K.: Unfoldomics of human diseases: linking protein intrinsic disorder with diseases. BMC Genomics 10(suppl. 1), S7 (2009), doi:1471-2164-10-S1-S7, [pii] 10.1186/1471-2164-10-S1-S7
15. Iakoucheva, L.M., Brown, C.J., Lawson, J.D., Obradovic, Z., Dunker, A.K.: Intrinsic disorder in cell-signaling and cancer-associated proteins. J. Mol. Biol. 323(3), 573–584 (2002), doi:S0022283602009695 [pii]
16. Pajkos, M., Meszaros, B., Simon, I., Dosztanyi, Z.: Is there a biological cost of protein disorder? Analysis of cancer-associated mutations. Mol. Biosyst. 8(1), 296–307 (2012), doi:10.1039/c1mb05246b
17. Cheng, Y., LeGall, T., Oldfield, C.J., Mueller, J.P., Van, Y.Y., Romero, P., Cortese, M.S., Uversky, V.N., Dunker, A.K.: Rational drug design via intrinsically disordered protein. Trends Biotechnol. 24(10), 435–442 (2006), doi:S0167-7799(06)00184-3, [pii] 10.1016/j.tibtech.2006.07.005
18. Metallo, S.J.: Intrinsically disordered proteins are potential drug targets. Curr. Opin. Chem. Biol. 14(4), 481–488 (2010),
doi:S1367-5931(10)00074-8, [pii] 10.1016/j.cbpa.2010.06.169
19. Uversky, V.N.: Natively unfolded proteins: a point where biology waits for physics. Protein Sci. 11(4), 739–756 (2002), doi:10.1110/ps.4210102
20. Dyson, H.J., Wright, P.E.: Coupling of folding and binding for unstructured proteins. Curr. Opin. Struct. Biol. 12(1), 54–60 (2002), doi:S0959440X02002890 [pii]
21. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E.: The Protein Data Bank. Nucleic Acids Res. 28(1), 235–242 (2000), doi:gkd090 [pii]
22. Gunasekaran, K., Tsai, C.J., Nussinov, R.: Analysis of ordered and disordered protein complexes reveals structural features discriminating between stable and unstable monomers. J. Mol. Biol. 341(5), 1327–1341 (2004), doi:10.1016/j.jmb.2004.07.002, [pii] S0022-2836(04)00801-0
23. Meszaros, B., Tompa, P., Simon, I., Dosztanyi, Z.: Molecular principles of the interactions of dis-ordered proteins. J. Mol. Biol. 372(2), 549–561 (2007), doi:S0022-2836(07)00920-5, [pii] 10.1016/j.jmb.2007.07.004

24. Uversky, V.N., Oldfield, C.J., Dunker, A.K.: Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling. J. Mol. Recognit. 18(5), 343–384 (2005), doi:10.1002/jmr.747
25. Dosztanyi, Z., Chen, J., Dunker, A.K., Simon, I., Tompa, P.: Disorder and sequence repeats in hub proteins and their implications for network evolution. J. Proteome Res. 5(11), 2985–2995 (2006), doi:10.1021/pr060171o
26. Bracken, C., Iakoucheva, L.M., Romero, P.R., Dunker, A.K.: Combining prediction, computation and experiment for the characterization of protein disorder. Curr. Opin. Struct. Biol. 14(5), 570–576 (2004), doi:S0959-440X(04)00137-X, [pii] 10.1016/j.sbi.2004.08.003
27. Garner, E., Cannon, P., Romero, P., Obradovic, Z., Dunker, A.K.: Predicting Disordered Regions from Amino Acid Sequence: Common Themes Despite Differing Structural Characterization. In: Genome Inform. Ser. Workshop Genome Inform., vol. 9, pp. 201–213 (1998)
28. Li, X., Romero, P., Rani, M., Dunker, A.K., Obradovic, Z.: Predicting Protein Disorder for N-, C-, and Internal Regions. In: Genome Inform. Ser. Workshop Genome Inform., vol. 10, pp. 30–40 (1999)
29. Radivojac, P., Obradovic, Z., Smith, D.K., Zhu, G., Vucetic, S., Brown, C.J., Lawson, J.D., Dunker, A.K.: Protein flexibility and intrinsic disorder. Protein Sci. 13(1), 71–80 (2004), doi:10.1110/ps.03128904
30. He, B., Wang, K., Liu, Y., Xue, B., Uversky, V.N., Dunker, A.K.: Predicting intrinsic disorder in proteins: an overview. Cell Res. 19(8), 929–949 (2009), doi:cr200987, [pii] 10.1038/cr.2009.87
31. Wootton, J.C.: Non-globular domains in protein sequences: automated segmentation using complexity measures. Comput. Chem. 18(3), 269–285 (1994), doi:0097-8485(94)85023-2 [pii]
32. Wootton, J.C., Federhen, S.: Analysis of compositionally biased regions in sequence databases. Methods Enzymol. 266, 554–571 (1996)
33. Romero, P., Obradovic, Z., Li, X., Garner, E.C., Brown, C.J., Dunker, A.K.: Sequence complexity of disordered protein. Proteins 42(1), 38–48 (2001), doi:10.1002/1097-0134(20010101)42:1<38::AID-PROT50>3.0.CO;2-3 [pii]
34. Vucetic, S., Obradovic, Z., Vacic, V., Radivojac, P., Peng, K., Iakoucheva, L.M., Cortese, M.S., Lawson, J.D., Brown, C.J., Sikes, J.G., Newton, C.D., Dunker, A.K.: DisProt: a database of protein disorder. Bioinformatics 21(1), 137–140 (2005), doi:10.1093/bioinformatics/bth476bth476 [pii]
35. Dutta, S., Burkhardt, K., Young, J., Swaminathan, G.J., Matsuura, T., Henrick, K., Nakamura, H., Berman, H.M.: Data deposition and annotation at the worldwide protein data bank. Mol. Biotechnol. 42(1), 1–13 (2009), doi:10.1007/s12033-008-9127-7
36. Mohan, A., Uversky, V.N., Radivojac, P.: Influence of sequence changes and environment on intrinsically disordered proteins. PLoS Comput. Biol., e1000497 (2009), doi:10.1371/journal.pcbi.1000497
37. De Biasio, A., Guarnaccia, C., Popovic, M., Uversky, V.N., Pintar, A., Pongor, S.: 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. 7(6), 2496–2506 (2008), doi:10.1021/pr800063u
38. Uversky, V.N., Gillespie, J.R., Fink, A.L.: Why are "natively unfolded" proteins unstructured under physiologic conditions? Proteins 41(3), 415–427 (2000), doi:10.1002/1097-0134(20001115)41:3<415::AID-PROT130>3.0.CO;2-7 [pii]

39. Galzitskaya, O.V., Garbuzynskiy, S.O., Lobanov, M.Y.: FoldUnfold: web server for the prediction of disordered regions in protein chain. Bioinformatics 22(23), 2948–2949 (2006), doi:btl504, [pii] 10.1093/bioinformatics/btl504
40. Xie, Q., Arnold, G.E., Romero, P., Obradovic, Z., Garner, E., Dunker, A.K.: The Sequence Attribute Method for Determining Relationships Between Sequence and Protein Disorder. In: Genome Inform. Ser. Workshop Genome Inform., vol. 9, pp. 193–200 (1998)
41. Campen, A., Williams, R.M., Brown, C.J., Meng, J., Uversky, V.N., Dunker, A.K.: TOP-IDP-scale: a new amino acid scale measuring propensity for intrinsic disorder. Protein Pept. Lett. 15(9), 956–963 (2008)
42. Linding, R., Russell, R.B., Neduva, V., Gibson, T.J.: GlobPlot: Exploring protein sequences for globularity and disorder. Nucleic Acids Res. 31(13), 3701–3708 (2003)
43. Cheng, J., Sweredoski, M., Baldi, P.: Accurate prediction of protein disordered regions by mining protein structure. Data Mining and Klowledge Discovery 11, 213–222 (2005)
44. Su, C.T., Chen, C.Y., Hsu, C.M.: iPDA: integrated protein disorder analyzer. Nucleic Acids Res. 35(Web Server Issue), W465–W472 (2007), doi:gkm353, [pii] 10.1093/nar/gkm353
45. Fuxreiter, M., Simon, I., Friedrich, P., Tompa, P.: Preformed structural elements feature in partner recognition by intrinsically unstructured proteins. J. Mol. Biol. 338(5), 1015–1026 (2004), doi:10.1016/j.jmb.2004.03.017, [pii] S0022283604003079
46. Suveges, D., Gaspari, Z., Toth, G., Nyitray, L.: Charged single alpha-helix: a versatile protein structural motif. Proteins 74(4), 905–916 (2009), doi:10.1002/prot.22183
47. Brown, C.J., Takayama, S., Campen, A.M., Vise, P., Marshall, T.W., Oldfield, C.J., Williams, C.J., Dunker, A.K.: Evolutionary rate heterogeneity in proteins with long disordered regions. J. Mol. Evol. 55(1), 104–110 (2002), doi:10.1007/s00239-001-2309-6
48. Daughdrill, G.W., Narayanaswami, P., Gilmore, S.H., Belczyk, A., Brown, C.J.: Dynamic behavior of an intrinsically unstructured linker domain is conserved in the face of negligible amino acid sequence conservation. J. Mol. Evol. 65(3), 277–288 (2007), doi:10.1007/s00239-007-9011-2
49. Peng, K., Radivojac, P., Vucetic, S., Dunker, A.K., Obradovic, Z.: Length-dependent prediction of protein intrinsic disorder. BMC Bioinformatics 7, 208 (2006), doi:1471-2105-7-208, [pii] 10.1186/1471-2105-7-208
50. Melamud, E., Moult, J.: Evaluation of disorder predictions in CASP5. Proteins 53(suppl. 6), 561–565 (2003), doi:10.1002/prot.10533
51. Jin, Y., Dunbrack Jr., R.L.: Assessment of disorder predictions in CASP6. Proteins 61(suppl. 7), 167–175 (2005), doi:10.1002/prot.20734
52. Bordoli, L., Kiefer, F., Schwede, T.: Assessment of disorder predictions in CASP7. Proteins 69(suppl. 8), 129–136 (2007), doi:10.1002/prot.21671
53. Noivirt-Brik, O., Prilusky, J., Sussman, J.L.: Assessment of disorder predictions in CASP8. Proteins 77(suppl. 9), 210–216 (2009), doi:10.1002/prot.22586
54. Monastyrskyy, B., Fidelis, K., Moult, J., Tramontano, A., Kryshtafovych, A.: Evaluation of disorder predictions in CASP9. Proteins 79(suppl. 10), 107–118 (2011), doi:10.1002/prot.23161
55. Dosztanyi, Z., Sandor, M., Tompa, P., Simon, I.: Prediction of protein disorder at the domain level. Curr. Protein Pept. Sci. 8(2), 161–171 (2007)

56. Schlessinger, A., Punta, M., Yachdav, G., Kajan, L., Rost, B.: Improved disorder prediction by combination of orthogonal approaches. PLoS One 4(2), e4433 (2009), doi:10.1371/journal.pone.0004433
57. 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 23(16), 2046–2053 (2007), doi:btm302, [pii] 10.1093/bioinformatics/btm302
58. Dosztanyi, Z., Meszaros, B., Simon, I.: Bioinformatical approaches to characterize intrinsically disordered/unstructured proteins. Brief Bioinform. 11(2), 225–243 (2010), doi:bbp061, [pii] 10.1093/bib/bbp061
59. Romero, Obradovic, Dunker, K.: Sequence Data Analysis for Long Disordered Regions Prediction in the Calcineurin Family. In: Genome Inform. Ser. Workshop Genome Inform., vol. 8, pp. 110–124 (1997)
60. Oldfield, C.J., Cheng, Y., Cortese, M.S., Romero, P., Uversky, V.N., Dunker, A.K.: Coupled folding and binding with alpha-helix-forming molecular recognition elements. Biochemistry 44(37), 12454–12470 (2005), doi:10.1021/bi050736e
61. Cheng, Y., Oldfield, C.J., Meng, J., Romero, P., Uversky, V.N., Dunker, A.K.: Mining alpha-helix-forming molecular recognition features with cross species sequence alignments. Biochemistry 46(47), 13468–13477 (2007), doi:10.1021/bi7012273
62. Radivojac, P., Obradovic, Z., Brown, C.J., Dunker, A.K.: Prediction of boundaries between intrinsically ordered and disordered protein regions. In: Pac. Symp. Biocomput., pp. 216–227 (2003)
63. Linding, R., Jensen, L.J., Diella, F., Bork, P., Gibson, T.J., Russell, R.B.: Protein disorder prediction: implications for structural proteomics. Structure 11(11), 1453–1459 (2003), doi:S0969212603002351 [pii]
64. Su, C.T., Chen, C.Y., Ou, Y.Y.: Protein disorder prediction by condensed PSSM considering propensity for order or disorder. BMC Bioinformatics 7, 319 (2006), doi:1471-2105-7-319, [pii] 10.1186/1471-2105-7-319
65. Schaffer, A.A., Aravind, L., Madden, T.L., Shavirin, S., Spouge, J.L., Wolf, Y.I., Koonin, E.V., Altschul, S.F.: Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res. 29(14), 2994–3005 (2001)
66. Ishida, T., Kinoshita, K.: PrDOS: prediction of disordered protein regions from amino acid sequence. Nucleic Acids Res. 35(Web Server Issue), W460–W464 (2007), doi:gkm363, [pii] 10.1093/nar/gkm363
67. Wang, L., Sauer, U.H.: OnD-CRF: predicting order and disorder in proteins using [corrected] conditional random fields. Bioinformatics 24(11), 1401–1402 (2008), doi:btn132, [pii] 10.1093/bioinformatics/btn132
68. MacCallum, R.: `http://www.sbc.su.se/~maccallr/disorder/` (date last accessed July 3, 2012)
69. Obradovic, Z., Peng, K., Vucetic, S., Radivojac, P., Dunker, A.K.: Exploiting heterogeneous sequence properties improves prediction of protein disorder. Proteins 61(suppl. 7), 176–182 (2005), doi:10.1002/prot.20735
70. Hirose, S., Shimizu, K., Inoue, N., Kanai, S., Noguchi, T.: Disordered region prediction by integrating POODLE series. In: CASP8 Proceedings 2008, pp. 14–15 (2008)
71. Bujnicki, J.M., Elofsson, A., Fischer, D., Rychlewski, L.: LiveBench-2: large-scale automated evaluation of protein structure prediction servers. Proteins (suppl. 5), 184–191 (2001), doi:10.1002/prot.10039 [pii]

72. Yang, Z.R., Thomson, R., McNeil, P., Esnouf, R.M.: RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. Bioinformatics 21(16), 3369–3376 (2005), doi:bti534, [pii] 10.1093/bioinformatics/bti534
73. Dosztanyi, Z., Csizmok, V., Tompa, P., Simon, I.: The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. J. Mol. Biol. 347(4), 827–839 (2005), doi:S0022-2836(05)00129-4, [pii] 10.1016/j.jmb.2005.01.071
74. Prilusky, J., Felder, C.E., Zeev-Ben-Mordehai, T., Rydberg, E.H., Man, O., Beckmann, J.S., Silman, I., Suss-man, J.L.: FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically un-folded. Bioinformatics 21(16), 3435–3438 (2005), doi:bti537, [pii] 10.1093/bioinformatics/bti537
75. Thomas, P.D., Dill, K.A.: An iterative method for extracting energy-like quantities from protein structures. Proc. Natl. Acad. Sci. U S A 93(21), 11628–11633 (1996)
76. Shortle, D.: Propensities, probabilities, and the Boltzmann hypothesis. Protein Sci. 12(6), 1298–1302 (2003), doi:10.1110/ps.0306903
77. 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 21(16), 3433–3434 (2005), doi:bti541, [pii] 10.1093/bioinformatics/bti541
78. Dosztanyi, Z., Meszaros, B., Simon, I.: ANCHOR: web server for predicting protein binding regions in disordered proteins. Bioinformatics 25(20), 2745–2746 (2009), doi:btp518, [pii] 10.1093/bioinformatics/btp518
79. Diella, F., Haslam, N., Chica, C., Budd, A., Michael, S., Brown, N.P., Trave, G., Gibson, T.J.: Understanding eukaryotic linear motifs and their role in cell signaling and regulation. Front Biosci. 13, 6580–6603 (2008), doi:3175 [pii]
80. Sigrist, C.J., Cerutti, L., Hulo, N., Gattiker, A., Falquet, L., Pagni, M., Bairoch, A., Bucher, P.: PROSITE: a documented database using patterns and profiles as motif descriptors. Brief Bioinform. 3(3), 265–274 (2002)
81. Neduva, V., Russell, R.B.: Linear motifs: evolutionary interaction switches. FEBS Lett. 579(15), 3342–3345 (2005), doi:S0014-5793(05)00461-8, [pii] 10.1016/j.febslet.2005.04.005
82. Stein, A., Aloy, P.: Contextual specificity in peptide-mediated protein interactions. PLoS One 3(7), e2524 (2008), doi:10.1371/journal.pone.0002524
83. Dinkel, H., Michael, S., Weatheritt, R.J., Davey, N.E., Van Roey, K., Altenberg, B., Toedt, G., Uyar, B., Seiler, M., Budd, A., Jodicke, L., Dammert, M.A., Schroeter, C., Hammer, M., Schmidt, T., Jehl, P., McGuigan, C., Dymecka, M., Chica, C., Luck, K., Via, A., Chatr-Aryamontri, A., Haslam, N., Grebnev, G., Edwards, R.J., Steinmetz, M.O., Meiselbach, H., Diella, F., Gibson, T.J.: ELM–the database of eukaryotic linear motifs. Nucleic Acids Res. 40(Database Issue), D242–D251 (2012), doi:gkr1064, [pii] 10.1093/nar/gkr1064
84. Davey, N.E., Trave, G., Gibson, T.J.: How viruses hijack cell regulation. Trends Biochem. Sci. 36(3), 159–169 (2011), doi:S0968-0004(10)00200-8, [pii] 10.1016/j.tibs.2010.10.002
85. Davey, N.E., Edwards, R.J., Shields, D.C.: Estimation and efficient computation of the true probability of recurrence of short linear protein sequence motifs in unrelated proteins. BMC Bioinformatics 11, 14 (2010), doi:1471-2105-11-14, [pii] 10.1186/1471-2105-11-14

86. Gibson, T.J.: Cell regulation: determined to signal discrete cooperation. Trends Biochem. Sci. 34(10), 471–482 (2009), doi:S0968-0004(09)00142-X, [pii] 10.1016/j.tibs.2009.06.007
87. Stein, A., Pache, R.A., Bernado, P., Pons, M., Aloy, P.: Dynamic interactions of proteins in complex networks: a more structured view. FEBS J. 276(19), 5390–5405 (2009), doi:EJB7251, [pii] 10.1111/j.1742-4658.2009.07251.x
88. Weatheritt, R.J., Luck, K., Petsalaki, E., Davey, N.E., Gibson, T.J.: The identification of short linear motif-mediated interfaces within the human interactome. Bioinformatics 28(7), 976–982 (2012), doi:bts072, [pii] 10.1093/bioinformatics/bts072
89. Lupas, A., Van Dyke, M., Stock, J.: Predicting coiled coils from protein sequences. Science 252(5009), 1162–1164 (1991), doi:252/5009/1162, [pii] 10.1126/science.252.5009.1162
90. Jones, D.T.: Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 292(2), 195–202 (1999), doi:10.1006/jmbi.1999.3091, [pii] S0022-2836(99)93091-7