BIOINFORMATICS REVIEW
Vol. 27 no. 11 2011, pages 1449–1454
doi:10.1093/bioinformatics/btr175

Structural bioinformatics
Advance Access publication April 14, 2011

Proteins without 3D structure: definition, detection and beyond

Ferenc Orosz* and Judit Ovádi*
Institute of Enzymology, Biological Research Center, Hungarian Academy of Sciences, Karolina út 29, Budapest, H-1113 Hungary

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Predictions, and experiments to a lesser extent, following the decoding of the human genome showed that a significant fraction of gene products do not have well-defined 3D structures. While the presence of structured domains traditionally suggested function, it was not clear what the absence of structure implied. These and many other findings initiated the extensive theoretical and experimental research into these types of proteins, commonly known as intrinsically disordered proteins (IDPs). Crucial to understanding IDPs is the evaluation of structural predictors based on different principles and trained on various datasets, which is currently the subject of active research. The view is emerging that structural disorder can be considered as a separate structural category and not simply as absence of secondary and/or tertiary structure. IDPs perform essential functions and their improper functioning is responsible for human diseases such as neurodegenerative disorders.

Contact: orosz@enzim.hu; ovadi@enzim.hu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on February 3, 2011; revised on March 25, 2011; accepted on March 30, 2011

1 DEFINITION AND CLASSIFICATION

There is currently a lack of consensus regarding the definition, terminology or nomenclature of proteins without well-defined 3D structure in their native (functional) state. These proteins lack a stable equilibrium conformation but exist as dynamic ensembles within which atom positions exhibit extreme temporal fluctuations without specific equilibrium values (Uversky and Dunker, 2010). The intrinsically disordered proteins (IDPs), the most frequently used term, just as others do not always differentiate whether the whole or (a) significant segment(s) of the sequences of these proteins are without defined structures. The Database of Protein Disorder (*DisProt*) defines IDP as 'a protein that contains at least one experimentally determined disordered region' (Sickmeier *et al.*, 2007). There are proteins disordered in full length while others contain both ordered and disordered parts termed as intrinsically disordered regions (IDRs; Obradovic *et al.*, 2005). These proteins are often named also as IDPs; however, the name 'proteins with IDR(s)' appears to be more correct.

A number of terms have been used to indicate the disordered characteristics of these proteins which are as follows: natively denatured (Schweers *et al.*, 1994), natively unfolded (Weinreb *et al.*, 1996), intrinsically unfolded (Baskakov *et al.*, 1999), intrinsically unstructured (Wright and Dyson, 1999), intrinsically disordered (Dunker *et al.*, 2000), exceptionally flexible (Ahmed *et al.*, 2007), natively unstructured (Schlessinger *et al.*, 2007a), naturally flexible (Uversky *et al.*, 2009), etc. The most often used synonyms are intrinsically unstructured/disordered proteins (IUPs/IDPs), although the pioneers of this field (Dunker *et al.* 2008, Uversky *et al.*, 2009) consider IUPs as a subset of IDPs without hydrophobic core and significant amounts of stable secondary structure.

The central dogma of the new protein structure–function paradigm, the *protein trinity/quartet*, consisting of the folded (ordered) state, the molten globule and the random coil (Dunker *et al.*, 2001), plus the pre-molten globule as the fourth unique thermodynamic state (Uversky, 2002), is that *any of these states may be the native state*, which is relevant to the biological function of a protein. Accordingly, the three types of intrinsic disorder can be classified as the native coil, native pre-molten globule (both considered as intrinsically unstructured) and native molten globule (Uversky *et al.*, 2009). Native coils, characterized by *extended disorder*, arise from chains having repulsion arising from a net charge, and these proteins and regions resemble the more classical idealized random coil. Pre-molten globules have no well-defined tertiary structure, may contain regions with transient and small amount of secondary structure and 3-fold larger hydrodynamic volume as expected for the folded state. Molten globules (*collapsed disorder*) possess secondary structure and folding pattern similar to the folded state (Uversky, 2002), with loosened, i.e. molten, tertiary interactions and exhibit an increase in hydrodynamic volume of no more than 50% (Uversky, 2002; Wright and Dyson, 1999). This classification, which is slightly artificial, is useful from conceptual and practical point of view, e.g. for the development of specific disorder predictors. However, we emphasize that according to our concept, in agreement with recently published ones (Rauscher and Pomés, 2010; Xue *et al.*, 2009), there is *a continuum of these states* with various degrees of compactness due to the different amounts and distribution of secondary and tertiary structure. Nevertheless, it is important to remember that this view does not influence the fact that each of these states, not only the folded one, can occur as the native state.

Structural disorder can be considered as a separate structural category, and not merely as a lack of secondary and/or tertiary structure (Tompa and Kalmar, 2010). The length distribution of disorder in the human proteome is scale free (follows a power law), with many short regions and also a significant incidence of very long disordered regions. This is in sharp contrast with the length distribution of conventional secondary structural elements, which

*To whom correspondence should be addressed.

shows a length limit near 50 residues, however, is highly reminiscent of the distribution of tertiary structural units (domains) in proteins. This behavior was correlated with the direct functional involvement of disorder (Section 5), which place structural disorder as a unique 'structural' level between secondary and tertiary structures (Tompa and Kalmar, 2010). In line with this, the sequences of intrinsically disordered domains differ significantly from random sequences; the deviation is comparable with that found between ordered and disordered domains (Teraguchi *et al.*, 2010) suggesting a 'structure–function' relationship for disordered regions.

2 OCCURRENCE

Bioinformatics predictions (Section 4) suggest that intrinsic structural disorder is a widespread phenomenon, especially in eukaryotes, where conservative estimations suggest that 5–15% of proteins are disordered in full sequence (IDPs), and about 35–50% of proteins have at least one IDR (more than 30 residues) (Ward *et al.*, 2004). It was interpreted that more disorder is needed for signaling and coordination among the various organelles of eukaryotes being more complex than prokaryotes (Dunker and Obradovic, 2001). Indeed, in mammals, 75% of signaling proteins are predicted to contain long disordered regions (Dunker *et al.*, 2008). In general, the more complex the organism is the more frequent occurrence of disorder can be found. [An interesting exception is that some protist parasites have the highest prevalence of disorder (Feng *et al.*, 2006)]. The various genome wide *in silico* studies based on Gene Ontology annotations and Swiss-Prot functional keywords suggest that the biological processes involving IDPs are as follows (Tompa, 2009; Tompa *et al.*, 2006; Ward *et al.*, 2004; Xie *et al.*, 2007): (i) transcription and its regulation; (ii) signal transduction and cell-cycle regulation; (iii) functioning of nucleic acid containing organelle; (iv) mRNA processing and splicing; and (v) cytoskeleton organization. These results from genome-wide predictions of intrinsic disorder and the results from other bioinformatics studies drew attention to these proteins. However, prediction of disorder foreruns its experimental identification and only relatively few experimentally characterized examples are known. There is experimental evidence for the structural disorder of about 1100 regions within 500 proteins, which is collected in the DisProt database (Sickmeier *et al.*, 2007).

3 EXPERIMENTAL IDENTIFICATION

Disordered proteins can be identified and characterized by using wide arsenal of the experimental methodologies (Daughdrill *et al.*, 2005; Eliezer, 2009; Mittag and Forman-Kay, 2007; Receveur-Bréchot *et al.*, 2006; Uversky and Longhi, 2010). Nevertheless, the absence of a well-defined structure in disordered proteins complicates their investigation, since the determination of unique high-resolution structure of IDPs is frequently not attainable. Instead, the goal usually is to obtain experimental constraints on the ensemble of states, including the detection of residual secondary structure, transient long-range contacts and regions of restricted or enhanced mobility (Eliezer, 2007). There is neither time nor space to cover these important experimental results. Here, we give only a brief description of the techniques which are most successfully used for identification of IDPs and/or IDRs. Most of them provide information for the global structures and do not identify the specific disordered region(s) within the molecule. For a more detailed description of specific methods, please see Supplementary Material File 1.

4 PREDICTION

Since the first predictors of protein disorder were published (Li *et al.*, 1999; Romero *et al.*, 1997), almost 60 predictors have been developed so far. The properties and the advantages/disadvantages of these predictors are summarized in several papers (Dosztányi and Tompa, 2008; Dosztányi *et al.*, 2010; Feng *et al.*, 2006, Ferron *et al.*, 2006; He *et al.*, 2009; Tompa, 2009; Uversky and Dunker, 2010). A very comprehensive recent review has been published by He *et al.* (2009), which contains detailed descriptions of the most often used predictors and references to the publicly available ones. The majority of the programs are accessible via public servers; links to many of them can be found in the Disordered Protein Database (http://www.DisProt.org) (Sickmeier *et al.*, 2007). Here, we shortly introduce merely some representative and frequently used disorder predictors (cf. also Supplementary Material File 2).

The predictors are based on different principles and can be classified into three main categories (Csizmók and Tompa, 2009; Tompa, 2009): (i) propensity-based predictors; (ii) machine learning algorithms; and (iii) algorithms based on interresidue contacts. These categories are not absolute since some of the methods use more than one of these features. Moreover, combined meta-servers also exist.

4.1 Propensity-based predictors

IDPs are significantly depleted in so-called order-promoting residues, including bulky hydrophobic (Ile, Leu and Val) and aromatic amino acid residues (Trp, Tyr and Phe), which would normally form the hydrophobic core of a globular protein, as well as Cys and Asn. On the other hand, there are so-called disorder-promoting amino acids, namely, Ala, Arg, Gly, Gln, Ser, Pro, Glu and Lys, which are substantially overrepresented in IDPs (Dunker *et al.*, 2001; Romero *et al.*, 2001). This specific amino acid composition is usually indicative for disorder. This propensity of IDPs was used to develop sophisticated prediction methods as well (Section 4.2).

Several methods are based on simple amino acid propensity scales. Their advantage is that they are easy to calculate and to interpret; however, they are limited to a single property. For example, due to their amino acid composition, low overall mean hydropathy and high mean net charge represent a unique structural feature of IDPs/IDRs (Uversky *et al.*, 2000). The mean hydropathy is defined as the sum of the hydropathies of all residues divided by the number of residues in the polypeptide. The mean net charge is the net charge, at pH 7.0, divided by the total number of residues. A plot of mean net charge versus mean hydropathy (the CH plot or Uversky plot) separates ordered and disordered proteins into distinct regions (Uversky, 2002). By calculating the distribution of these features for a pre-defined sequence window, Prilusky *et al.* (2005) used this idea to design a per-residue disorder predictor, *FoldIndex*. Another predictor, the *GlobPlot* algorithm (Linding *et al.*, 2003a), uses the relative propensity of amino acid residues to be in an ordered or disordered state applying an amino acid scale based on the difference in the probability for a given amino acid to be in regular secondary structure or to be in random coil. The basic algorithm behind GlobPlot is simple and very fast, representing a sum function.

4.2 Machine learning algorithms

The prediction of protein disorder can be viewed as a classic binary classification problem and can be addressed by standard machine learning techniques as *artificial neural networks* (NNs) and *support vector machines* (SVMs). The majority of the methods developed belong to these categories. They are trained on datasets of disorder and order and evaluate intrinsic disorder on a per-residue basis. Their underlying assumption is that sequence features calculated from a local sequence window can be directly mapped into the property of order or disorder.

The Dunker's lab was the pioneer of predictors (Romero *et al.*, 1997); then their *PONDR*® family of algorithms has been continuously developed and improved (Li *et al.*, 1999; Obradovic *et al.*, 2005; Peng *et al.*, 2005, 2006; Romero *et al.*, 1997, 2001). They typically use NNs, although in a few cases SVMs are also included. PONDR® algorithms based on the fact that the amino acid composition in a window of N amino acids for ordered proteins are distinguishable from the composition for disordered proteins. Beside merely amino acid composition itself, they use as inputs some attributes derived from composition as well. These various types of attributes are weighted and combined in a non-linear manner. The training datasets are different in the various family members (missing residues of X-ray structures; variously characterized long disordered regions; DisProt). Accordingly, there are predictors for short and long (>30 amino acid) regions, and for N- and C-terminal and internal ones. For their detailed descriptions, see the authors' recent review (He *et al.*, 2009) and the group's home page (http://www.pondr.com).

DisEMBL designed by Linding *et al.* (2003b) consists of three separate NN predictors, to predict three kinds of disordered structures in proteins, which represent residues within 'loops/coils', 'hot loops (loops with high B-factors, i.e. with high degree of mobility)' or those that are missing from the PDB X-ray structures. Thus, it performs better on short disordered regions.

Another NN algorithm, *RONN*, developed by Yang *et al.* (2005) is based on 'functional alignments'. The main idea is that if two proteins have similar biological functions, in this case the similar tendencies to be ordered/disordered, then their sequences are also similar. In the training process, the similarity of sequences is evaluated by sequence alignment techniques using a mutation matrix to score the similarity. These scores of sequence alignments are then used for training.

The most often used SVM method is *DISOPRED2* (Ward *et al.*, 2004) where the input data are generated by sequence alignment using PSI-BLAST, and which is trained on a database of amino acids missing from PDB structures. Thus, the prediction is better on short disordered regions in the context of globally ordered proteins. The fact that the database contain much more ordered than disordered residues (176 550 versus 4590) is balanced by formulating the SVM to place greater cost of misclassification for points from the minority (disordered) class than from the majority (ordered) class. This is the reason for the low false positivity of DISOPRED2. Compared with other disorder predictors, the main difference is that DISOPRED2 is directly trained on the whole sequence rather than measures of amino acid composition, sequence complexity or biophysical properties. Prediction accuracy of this method depends on the number of homologs used for sequence alignment.

Additional methods which apply a second level of prediction using the output of the first level prediction as an input are the *POODLE* algorithms. They employ SVMs with radial basis kernels for training; the input is constructed from physico-chemical properties using PSI-BLAST profiles. POODLE-S (Shimizu *et al.*, 2007) and POODLE-L (Hirose *et al.*, 2007), which aim to predict disordered segments shorter and longer than 40 residues, respectively, calculate the input vector by using physico-chemical features and a reduced amino acid set of position-specific scoring matrices or from hydropathy, average contact density propensity, mean net charge, sequence complexity, amino acid compositions relative to the composition of disordered and ordered training sets and secondary structure preferences.

4.3 Prediction methods based on interresidue contacts

Limitations due to the biased and insufficient databases can be overcome by the methods based on structural and energetic considerations, which do not rely on experimental data on protein disorder. The prominent representatives of these methods are *FoldUnfold* (Galzitskaya *et al.*, 2006; Garbuzynskiy *et al.*, 2004), *IUPred* (Dosztányi *et al.*, 2005a, b), and *Ucon* (Schlessinger *et al.*, 2007a). The main idea of these methods is that the disorder of proteins is originated from the lack or low level of the *interresidue contacts* which cannot compensate the large decrease in conformational entropy during folding (Tompa, 2009). The importance of interresidue contacts, especially that of the heavily interacting residue clusters (stabilization centers) is essential in the maintenance of the folded protein structure (Dosztányi *et al.*, 1997). Intuitively, it can be thought that the lack of them favors protein disorder, as it was found indeed in several cases (Orosz *et al.*, 2004).

FoldUnfold based on the statistical analysis of residue contact numbers. The summation of the interresidue contact numbers of the amino acids of a protein is indicative for its folded/unfolded character. Two residues are considered in contact if any pair of their heavy atoms is within 8.0 Å to each other. To express the average contact number of residues within a given distance in a protein structure, the mean packing density of residues is calculated. It was demonstrated that regions with low-expected packing density correspond to the disordered segments.

Ucon combines a former predictor, PROFcon, for long-range protein-specific internal contacts (Punta and Rost, 2005) with a generic pairwise potential to predict unstructured regions longer than 30 amino acids. It combines information from alignments, from predictions of secondary structure and solvent accessibility, from the region between two residues and from the average properties of the entire protein.

The core of *IUPred* is a method that renders the direct estimation of the interaction energies using exclusively the protein sequence possible. In this approach, the estimated energy for each residue depends on the amino acid type and on the amino acid composition of the sequential neighborhood. Generally, residues with less favorable predicted energies are more likely to be disordered. The parameters of this method are derived exclusively 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 stabilizes the method substantially if compared with methods where a large number of parameters are trained on a limited and sometimes biased disordered protein dataset. IUPred performs comparatively well for predicting long disordered segments, and has a good sensitivity, i.e. does not miss a significant number of disordered residues.

However, a problem arises by the presence of conserved cysteines and/or of metal-binding motifs which can cause uncertain local predictions of disorder within these regions using the methods based on interresidue contacts. These predictors may display features typifying disorder while the protein region gains structure upon disulfide formation or binding to metal ions (Ferron *et al.*, 2006). This problem can in part be handled using methods predicting metal-binding sites and disulfide bridges of proteins from their sequence (Lippi *et al.*, 2008).

4.4 Metapredictors

Generally, it is a good idea not to rely on one single algorithm when predicting disorder. Instead, as these algorithms all capture different aspects of the structural properties of proteins, they can complement each other to give a more complete picture. Recently, a new direction in the development of disorder predictors based on the creation of metapredictors has attracted attention. These metapredictors (*metaPrDOS, MeDor, MD* and *PONDR-FIT*) combine the outputs of several individual predictors (Ishida and Kinoshita, 2008; Lieutaud *et al.*, 2008; Schlessinger *et al.*, 2009; Xue *et al.*, 2010a). They can be applied either at the residue level or at the whole sequence level. The individual predictors constituting metapredictors are based on different philosophies, the strength and weakness of which can be balanced by their combination. MetaPrDOS (Ishida and Kinoshita, 2008) uses SVM to integrate residue-level predictions from several algorithms and was trained on a group of PDB-extracted proteins that all have regions of missing electron density in their crystal structures, and the sequence identities among these proteins are <20%. Meta-Disorder predictor (MD) (Schlessinger *et al.*, 2009) uses NN and the training datasets were proteins from PDB and DisProt. The metapredictors improved the prediction accuracies which were several percentage points higher (max. 10 %) on various datasets in comparison with the values estimated for the individual predictors.

4.5 Future directions

The performance of disorder predictors has been compared in the CASP (Critical Assessment of Structure Prediction) experiments (Bordoli *et al.*, 2007; Jin and Dunbrack, 2005; Melamud and Moult, 2003; Noivirt-Brik *et al.*, 2009). However, these comparisons are considered to be rather biased since 'the performance of the methods depends on both the type of disorder and evaluation criteria' (Tompa, 2009) as discussed in details in several other papers as well (Dosztányi *et al.*, 2010; Schlessinger *et al.*, 2007b). Moreover, the predictors focus on different type ('flavors') of disorder, thus predictors trained on disorder of one type of protein often achieve poor accuracy on disorder of proteins of a different type, as recognized already at the advent of IDP research (Vucetic *et al.*, 2003).

Currently, the per-residue prediction accuracies of these methods have risen to about 80% (Dunker *et al.*, 2008). The limitation to further improvement comes from inaccuracy in the ordered and disordered protein data (Dosztányi *et al.*, 2010; He *et al.*, 2009). The performance of disorder prediction methods critically depends on the dataset used for testing and the type of disorder (e.g. extended or collapsed) studied (Dosztányi *et al.*, 2010; Schlessinger *et al.*, 2007b; Vucetic *et al.*, 2003). Datasets of experimentally verified ordered and disordered regions contain many mis-classified segments; moreover, the latter ones are not sufficiently large for prediction of very high level of accuracy. Various datasets of disordered protein sequences exhibit variations in their sequential bias. Differences can be observed depending also on the experimental method used for identification of the disordered regions (Dosztányi *et al.*, 2010), on their length, and on the location in the sequence (N- and C-terminal, middle regions) (Li *et al.*, 1999). Although these differences are smaller compared with the differences observed between ordered and disordered proteins, they should be taken into account during the development of prediction methods. Thus, it was suggested that predictors that go beyond the binary classification of proteins as ordered or disordered are necessary (Dosztányi *et al.*, 2010; He *et al.*, 2009).

5 FUNCTION

Disordered proteins can be separated into two main functional classes, based on their *in vivo* activities: *entropic chains* and IDPs involved in *molecular recognition* (Tompa 2002, 2005). There are IDPs which do not have a folded or ordered state under any known conditions, while others are capable of folding under certain circumstances, i.e. upon binding to a partner, termed as '*non-folders*' and '*folders*', respectively (Rauscher and Pomés, 2010). Entropic chains are necessarily 'non-folders', since their functions rely on their high-conformational entropy: their functions are derived by populating many accessible conformations without well-defined folded structure (Tompa, 2009). On the contrary, IDPs involved in molecular recognition due to their interacting potencies are generally 'folders' that become (partly) ordered upon binding to their targets. The binding can be permanent (*scavengers, effectors* and *assemblers*) or transient (*display sites* and *chaperones*) (Tompa and Kovacs, 2010). Scavengers and assemblers usually bind to multiple partners. Scavengers store and neutralize small molecules, while assemblers support the assembly of multi-protein complexes. Effectors regulate the activity of partner proteins. Display sites expose sites for post-translational modifications such as phosphorylation or limited proteolysis, whereas chaperones bind to the partner molecule to facilitate its correct folding preventing its aggregation or proteolysis.

5.1 Binding to partners

The folding of IDPs during molecular recognition is analogous to protein folding of globular proteins, since both processes involve a thermodynamically stable folded state and an unfolded state of higher conformational entropy (Verkhivker *et al.*, 2003, 2005). Since many IDPs and IDRs fold upon binding to their targets (Wright and Dyson, 2009), a challenging question is whether folding occurs before binding or binding occurs before folding? The two extremes are induced folding and conformational selection (Wright and Dyson, 2009). In the case of the first mechanism, the protein associates with its binding partner in a disordered state and subsequently folds in association with the target protein. In the conformational selection mechanism, the target protein 'selects' a conformation closely approximating that of the bound form from the ensemble of conformations populated by the IDP when free in solution. This question is analogous to and generalization of the induced fit—fluctuation fit (conformational selection) duality of molecular recognition (Vértessy and Orosz, 2011). In real systems, one or another or both mechanism(s) can be favored (Wright and Dyson, 2009).

In general, the binding of IDPs differs from that of ordered proteins since they often bind their partner via short recognition elements [*MoRFs* (*mo*lecular *r*ecognition *f* eatures)] (Fuxreiter *et al.*, 2007; Mohan *et al.*, 2006; Oldfield *et al.*, 2005) in a structurally adaptive process termed *disorder-to-order transition* (Verkhivker *et al.*, 2003). Structural disorder may confer significant functional advantages for IDPs, such as rapid binding to the partner molecule, the combination of high specificity with weak and reversible interaction and the ability to carry out more than one function either via multiple interaction sites or through regions specific to distinct partners (Tompa, 2002, 2005; Tompa *et al.*, 2005). They can fold into different structures on binding to different target proteins (‘*one-to-many binding mode*’), with different functional outcomes (functional promiscuity or moonlighting) (Tompa *et al.*, 2005). In contrast to this, the ‘*many-to-one binding mode*’ was also suggested, when many different IDPs may bind alternatively to one site on a single ordered partner, by which different IDPs fold into similar conformations (Uversky *et al.*, 2009). It was suggested that these binding features of IDPs may explain their ‘organizing role’ in protein–protein interaction networks as so-called ‘hub’ proteins (Dunker *et al.*, 2008).

Some binding sites can be found as linear motifs (LMs; Puntervoll *et al.*, 2003), short segments involved in the molecular recognition of proteins. It was shown that there is a connection between LMs and molecular recognition elements of IDPs. LMs are embedded in locally unstructured/highly flexible regions, while their amino acid composition exhibits a mixture characteristic of folded and disordered proteins (Fuxreiter *et al.*, 2007).

As we discussed above, there are many algorithms for predicting IDPs; however, the methods for predicting regions undergoing disorder-to-order transition upon protein binding is rather limited. An algorithm proposed by the Dunker’s and Uversky’s labs, *α-MoRF-PredII*, combines two bioinformatic tools, sequence alignment and disorder prediction, to find possible binding partners in protein databases and identify the interaction sites. The method is based on the identification of the above-mentioned MoRFs, which are short segments expected to have a high propensity for folding upon binding and that are located within regions of disorder (Cheng *et al.*, 2007; Mohan *et al.*, 2006; Oldfield *et al.*, 2005). Very recently, the authors hypothesized that not only MoRFs with similar sequences can be aligned but also those of with reversed sequential order (‘retro-MoRFs’). Applying this theory, they developed a software package named *PONDR-RIBS*, which aligns protein segments, predicts disorder and interaction regions (Xue *et al.*, 2010b). However, experimental verification of this new method is needed.

A recent method, *ANCHOR*, based on the principles behind the IUPred algorithm (Section 4.3), has been developed for this aim (Mészáros *et al.*, 2009). The essential feature of these binding segments is that they exist in a disordered state in isolation, but they can favorably interact with a globular protein and adopt a rigid conformation upon binding. Based on this model, the combination of the high disordered tendency of the 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 (Dosztányi *et al.*, 2010).

For a discussion of the connection of intrinsic disorder and alternative splicing and diseases, respectively, please see Supplementary Material File 1.

Funding: Hungarian National Scientific Research Fund OTKA (T-067963 to J.O.); the European Commission (DCI-ALA/19.09.01/10/21526/245-297/ALFA 111(2010)29 to J.O.); the European Concerted Research Action (COST Action TD0905 to J.O.).

Conflict of Interest: none declared.

REFERENCES

Ahmed,M.A. *et al.* (2007) The BG21 isoform of Golli myelin basic protein is intrinsically disordered with a highly flexible amino-terminal domain. *Biochemistry*, **46**, 9700–9712.

Baskakov,I.V. *et al.* (1999) Trimethylamine N-oxide-induced cooperative folding of an intrinsically unfolded transcription-activating fragment of human glucocorticoid receptor. *J. Biol. Chem.*, **274**, 10693–10696.

Bordoli,L. *et al.* (2007) Assessment of disorder predictions in CASP7. *Proteins*, **69** (Suppl. 8), 129–136.

Cheng,Y. *et al.* (2007) Mining α-helix-forming molecular recognition features α-MoRFs with cross species sequence alignments. *Biochemistry*, **46**, 13468–13477.

Csizmók,V. and Tompa,P. (2009) Structural disorder and its connection with misfolding diseases. In Ovádi,J. and Orosz,F. (eds) *Protein Folding and Misfolding: Neurodegenerative Diseases*. Springer, pp. 1–19.

Daughdrill,G.W. *et al.* (2005) Natively disordered proteins. In Buchner,J. and Kiefhaber,T. (eds) *Protein Folding Handbook*. Wiley-VCH, Verlag GmbH & Co. KGaA, Weinheim, Germany, pp. 271–353.

Dosztányi,Z. and Tompa,P. (2008) Prediction of protein disorder. *Methods Mol. Biol.*, **426**, 103–115.

Dosztányi,Z. *et al.* (1997) Stabilization centers in proteins: identification, characterization and predictions. *J. Mol. Biol.*, **272**, 597–612.

Dosztányi,Z. *et al.* (2005a) The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J. Mol. Biol.*, **347**, 827–839.

Dosztányi,Z. *et al.* (2005b) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, **21**, 3433–3434.

Dosztányi,Z. *et al.* (2010) Bioinformatical approaches to characterize intrinsically disordered/unstructured proteins. *Brief. Bioinform.*, **11**, 225–243.

Dunker,A.K. and Obradovic,Z. (2001) The protein trinity – linking function and disorder. *Nat. Biotechnol.*, **19**, 805–806.

Dunker,A.K. *et al.* (2000) Intrinsic protein disorder in complete genomes. *Genome Inform. Ser. Workshop Genome Inform.*, **11**, 161–171.

Dunker,A.K. *et al.* (2001) Intrinsically disordered protein. *J. Mol. Graph. Model.*, **19**, 26–59.

Dunker,A.K. *et al.* (2008) The unfoldomics decade: an update on intrinsically disordered proteins. *BMC Genomics*, **9** (Suppl. 2):S1.

Eliezer,D. (2007) Characterizing residual structure in disordered protein States using nuclear magnetic resonance. *Methods Mol. Biol.*, **350**, 49–67.

Eliezer,D. (2009) Biophysical characterization of intrinsically disordered proteins. *Curr. Opin. Struct. Biol.*, **19**, 23–30.

Feng,Z.P. *et al.* (2006) Abundance of intrinsically unstructured proteins in *P. falciparum* and other apicomplexan parasite proteomes. *Mol. Biochem. Parasitol.*, **150**, 256–267.

Ferron,F. *et al.* (2006) A practical overview of protein disorder prediction methods. *Proteins*, **65**, 1–14.

Fuxreiter,M. *et al.* (2007) Local structural disorder imparts plasticity on linear motifs. *Bioinformatics*, **23**, 950–956.

Galzitskaya,O.V. *et al.* (2006) FoldUnfold: web server for the prediction of disordered regions in protein chain. *Bioinformatics*, **22**, 2948–2949.

Garbuzynskiy,S.O. *et al.* (2004) To be folded or to be unfolded. *Protein Sci.*, **13**, 2871–2877.

He,B. *et al.* (2009) Predicting intrinsic disorder in proteins: an overview. *Cell Res.*, **19**, 929–949.

Hirose,S. *et al.* (2007) POODLE-L: a two level SVM prediction system for reliably predicting long disordered regions. *Bioinformatics*, **23**, 2046–2053.

Ishida,T. and Kinoshita,K. (2008) Prediction of disordered regions in proteins based on the meta approach. *Bioinformatics*, **24**, 1344–1348.

Jin,Y. and Dunbrack,R.L. Jr (2005) Assessment of disorder predictions in CASP6. *Proteins*, **61** (Suppl. 7), 167–175.

Li,X. *et al.* (1999) Predicting protein disorder for N-, C-, and internal regions. *Genome Inform. Ser. Workshop Genome Inform.*, **10**, 30–40.

Lieutaud,P. *et al.* (2008) MeDor: a metaserver for predicting protein disorder. *BMC Genomics*, **9**, S25.

Linding,R. *et al.* (2003a) GlobPlot: exploring protein sequences for globularity and disorder. *Nucleic Acids Res.*, **31**, 3701–3708.

Linding,R. *et al.* (2003b) Protein disorder prediction: implications for structural proteomics. *Structure*, **11**, 1453–1459.

Lippi,M. *et al.* (2008) MetalDetector: a web server for predicting metal-binding sites and disulfide bridges in proteins from sequence. *Bioinformatics*, **24**, 2094–2095.

Melamud,E. and Moult,J. (2003) Evaluation of disorder predictions in CASP5. *Proteins*, **53** (Suppl. 6), 561–565.

Mittag,T. and Forman-Kay,J.D. (2007) Atomic-level characterization of disordered protein ensembles. *Curr. Opin. Struct. Biol.*, **17**, 3–14.

Mohan,A. *et al.* (2006) Analysis of molecular recognition features (MoRFs). *J. Mol. Biol.*, **362**, 1043–1059.

Mészáros,B. *et al.* (2009) Prediction of protein binding regions in disordered proteins. *PLoS Comput Biol.*, **5**, e1000376.

Noivirt-Brik,O. *et al.* (2009) Assessment of disorder predictions in CASP8. *Proteins*, **77** (Suppl. 9), 210–216.

Obradovic,Z. *et al.* (2005) Exploiting heterogeneous sequence properties improves prediction of protein disorder. *Proteins*, **61** (Suppl. 7), 176–182.

Oldfield,C.J. *et al.* (2005) Coupled folding and binding with alpha-helix-forming molecular recognition elements. *Biochemistry*, **44**, 12454–12470.

Orosz,F. *et al.* (2004) TPPP/p25: from unfolded protein to misfolding disease: prediction and experiments. *Biol. Cell*, **96**, 701–711.

Peng,K. *et al.* (2005) Optimizing long intrinsic disorder predictors with protein evolutionary information. *J. Bioinform. Comput. Biol.*, **3**, 35–60.

Peng,K. *et al.* (2006) Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics*, **7**, 208.

Prilusky,J, *et al.* (2005) FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics*, **21**, 3435–3438.

Punta,M. and Rost,B. (2005) PROFcon: novel prediction of long-range contacts. *Bioinformatics*, **21**, 2960–2968.

Puntervoll,P. *et al.* (2003) ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins, *Nucleic Acids Res.*, **31**, 3625–3630.

Rauscher,S. and Pomés,R. (2010) Molecular simulations of protein disorder. *Biochem. Cell Biol.*, **88**, 269–290.

Receveur-Bréchot,V. *et al.* (2006) Assessing protein disorder and induced folding. *Proteins*, **62**, 24–45.

Romero,P. *et al.* (1997) Identifying disordered regions in proteins from amino acid sequence. *Proc. IEEE Int. Conf. Neural Netw.*, **1**, 90–95.

Romero,P. *et al.* (2001) Sequence complexity of disordered protein. *Proteins*, **42**, 38–48.

Schlessinger,A. *et al.* (2007a) Natively unstructured regions in proteins identified from contact predictions. *Bioinformatics*, **23**, 2376–2384.

Schlessinger,A. *et al.* (2007b) Natively unstructured loops differ from other loops. *PLoS Comput. Biol.*, **3**, e140.

Schlessinger,A. *et al.* (2009) Improved disorder prediction by combination of orthogonal approaches. *PLoS One*, **4**, e4433.

Schweers,O. *et al.* (1994) Structural studies of tau protein and Alzheimer paired helical filaments show no evidence for beta-structure. *J. Biol. Chem.*, **269**, 24290–24297.

Sickmeier, M. *et al.* (2007) DisProt: the database of disordered proteins. *Nucleic Acids Res.*, **35**, D786–D793.

Shimizu,K. *et al.* (2007) POODLE-S: web application for predicting protein disorder by using physicochemical features and reduced amino acid set of a position-specific scoring matrix. *Bioinformatics*, **23**, 2337–2338.

Teraguchi,S. *et al.* (2010) Intrinsically disordered domains deviate significantly from random sequences in mammalian proteins. *BMC Bioinformatics*, **11** (Suppl 7), S7.

Tompa,P. (2002) Intrinsically unstructured proteins. *Trends Biochem. Sci.*, **27**, 527–533.

Tompa,P. (2005). The interplay between structure and function in intrinsically unstructured proteins. *FEBS Lett.*, **579**, 3346–3354.

Tompa,P. (2009) *Structure and Function of Intrinsically Disordered Proteins.* CRC Press/Taylor and Francis Group, Boca Raton, FL.

Tompa,P. and Kalmar,L. (2010) Power law distribution defines structural disorder as a structural element directly linked with function. *J. Mol. Biol.*, **403**, 346–350.

Tompa,P. and Kovacs,D. (2010) Intrinsically disordered chaperones in plants and animals. *Biochem. Cell Biol.*, **88**, 167–174.

Tompa,P. *et al.* (2005) Structural disorder throws new light on moonlighting. *Trends Biochem. Sci.*, **30**, 484–489.

Tompa,P. *et al.* (2006) Prevalent structural disorder in *E.coli* and *S.cerevisiae* proteomes. *J. Proteome Res.*, **5**, 1996–2000.

Uversky,V.N. (2002) Natively unfolded proteins: a point where biology waits for physics. *Protein Sci.*, **11**, 739–756.

Uversky,V.N. and Dunker,A.K. (2010) Understanding protein non-folding. *Biochim. Biophys. Acta*, **1804**, 1231–1264.

Uversky,V.N. and Longhi,S. (eds) (2010) Instrumental analysis of intrinsically disordered proteins: assessing structure and conformation. In Uversky,V.N. (series ed.) *The Wiley Series in Protein and Peptide Science*. John Wiley & Sons, Inc, Hoboken, NJ, USA.

Uversky,V.N. *et al.* (2000) Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins*, **41**, 415–427.

Uversky,V.N. *et al.* (2009) Unfoldomics of human diseases: linking protein intrinsic disorder with diseases. *BMC Genomics*, **10** (Suppl. 1), S7.

Verkhivker,G.M. *et al.* (2003) Simulating disorder-order transitions in molecular recognition of unstructured proteins: where folding meets binding. *Proc. Natl Acad. Sci. USA*, **100**, 5148–5153.

Verkhivker,G.M. (2005) Protein conformational transitions coupled to binding in molecular recognition of unstructured proteins. *Proteins*, **58**, 706–716.

Vértessy,B.G. and Orosz,F. (2011) From "fluctuation fit" to "conformational selection": evolution, rediscovery, and integration of a concept. *Bioessays*, **33**, 30–34.

Vucetic,S. *et al.* (2003) Flavors of protein disoder. *Proteins*, **52**, 573–584.

Ward,J.J. *et al.* (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.*, **337**, 635–645.

Weinreb,P.H. *et al.* (1996) NACP, a protein implicated in Alzheimer's disease and learning, is natively unfolded. *Biochemistry*, **35**, 13709–13715.

Wright,P.E. and Dyson,H.J. (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.*, **293**, 321–331.

Wright, P.E. and Dyson, H.J. (2009) Linking folding and binding. *Curr. Opin. Struct. Biol.*, **19**, 31–38.

Yang,Z.R. *et al.* (2005) RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics*, **21**, 3369–3376.

Xie,H. *et al.* (2007) Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions. *J. Proteome Res.*, **6**, 1882–1898.

Xue,B. *et al.* (2009) CDF it all: consensus prediction of intrinsically disordered proteins based on various cumulative distribution functions. *FEBS Lett.*, **583**, 1469–1474.

Xue,B. *et al.* (2010a) PONDR-FIT: A meta-predictor of intrinsically disordered amino acids. *Biochim. Biophys. Acta*, **1804**, 996–1010.

Xue,B. *et al.* (2010b) Retro-MoRFs: identifying protein binding sites by normal and reverse alignment and intrinsic disorder prediction. *Int. J. Mol. Sci.*, **11**, 3725–3747.