The intrinsically disordered proteins (IDPs), the most frequently
characteristics of these proteins which are as follows: natively
IDR(s)’ appears to be more correct.
are often named also as IDPs; however, the name ‘proteins with
disordered regions (IDRs; Obradovic
contain both ordered and disordered parts termed as intrinsically
one experimentally determined disordered region’ (Sickmeier
whole or (a) significant segment(s) of the sequences of these
used term, just as others do not always differentiate whether the
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 Kalmár, 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
These results from genome-wide predictions of intrinsic disorder show 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; Miltag and Forman-Kay, 2007; Receveur-Brechot 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 (Dosphzinai and Tompa, 2008; Doszhanyi 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 (Csizmok and Tompa, 2009; Tompa, 2009): (i) propensity-based predictors; (ii) machine learning algorithms; and (iii) algorithms based on inter residue 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, Glu, 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 hydrophathies 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...
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 POND® 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. POND® 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. Besides 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 homepage (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/tribs’, ‘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 hydrophathy, 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; Garbarszynski 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 (Oroso 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 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 Pomes, 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 (Véritessy 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 (molecular recognition features)) (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., 2003). This question is analogous to and generalization of promiscuity in proteins since they often bind their partner via short recognition 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-PredLI, 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 POND-ReRBS, 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 sites 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 (Doustájá 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/19.09.01/1021526245-297/ALFA 111/201029 to J.O.); the European Concerted Research Action (COST Action TD0905 to J.O.).

**Conflict of Interest:** none declared.

**REFERENCES**


Psiaki, J. et al. (2005) FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unstructured. Proteins, 58, 345–348.


