

Prediction of Intrinsic Disorder and Its Use in Functional Proteomics

Vladimir N. Uversky, Predrag Radivojac, Lilia M. Iakoucheva, Zoran Obradovic, and A. Keith Dunker

Summary

The number of experimentally verified, intrinsically disordered (ID) proteins is rapidly rising. Research is often focused on a structural characterization of a given protein, looking for several key features. However, ID proteins with their dynamic structures that interconvert on a number of time-scales are difficult targets for the majority of traditional biophysical and biochemical techniques. Structural and functional analyses of these proteins can be significantly aided by disorder predictions. The current advances in the prediction of ID proteins and the use of protein disorder prediction in the fields of molecular biology and bioinformatics are briefly overviewed herein. A method is provided to utilize intrinsic disorder knowledge to gain structural and functional information related to individual proteins, protein groups, families, classes, and even entire proteomes.

Key Words: Intrinsically disordered protein; natively unfolded protein; intrinsically unstructured protein; protein flexibility; disorder prediction; protein function.

1. Introduction

Although the protein sequence-structure-function paradigm (well known as the “lock-and-key” hypothesis [1]), according to which a protein can achieve its biological function only on folding into a unique, structured state determined by its amino acid sequence, was a dominating view for more than 100 yr, it is recognized now that the phenomenon of functional intrinsic disorder is highly abundant in nature. For example, only less than one-third of the crystal structures in the protein data bank (PDB) are completely devoid of disorder (2). In fact, recent discoveries of intrinsically disordered (ID) or natively unstructured proteins have significantly broadened the understanding of protein functionality and revealed a new and unexpected role of dynamics, plasticity, and flexibility

in protein function. Importantly, bioinformatics played a key role in transforming a set of anecdotal examples of intrinsically disordered proteins (IDPs), which were originally considered to be intriguing exceptions within the protein realm, into a very promising branch of protein science.

1.1. Defining and Identifying IDPs

IDPs or ID protein regions are those that fail to form specific three-dimensional (3D) structure under physiological conditions *in vitro*. They are also known as partially folded (3), flexible (4), mobile (5), rheomorphic (6), natively denatured (7), natively unfolded (8), intrinsically unstructured (9), ID (10), and natively disordered (11). Furthermore, several other names representing different combinations of “natively, naturally, and intrinsically” with “unfolded, unstructured, flexible, mobile, and denatured,” are present in literature (12). The interested reader will find the discussion of the etymology of the term “ID” in a recent review (13). In contrast to the ordered proteins, the atoms and dihedral angles of IDPs do not have equilibrium positions. Instead, IDPs exist as highly dynamic ensembles whose atoms and backbone Ramachandran angles fluctuate significantly over time. An ID region can be as short as a few amino acid residues, or it can propagate through the long disordered loops, ends, domains, or even through entire proteins (13). In the authors’ view, an IDP is a protein that contains at least one disordered region.

Functional ID regions exist in at least two different structural forms: molten globule-like (collapsed) and random coil-like (extended) (14). Later, the existence of another functional disordered form, the premolten globule, which appears to be a distinct category between extended and molten-globular conformations, was suggested (15). Thus, protein function might be associated with three (or four) distinct conformations: ordered, molten globule, (premolten globule), and random coil, and with the transitions between them. These hypotheses are known as the protein-trinity (14) or protein-quartet models (15).

IDPs can be identified by the variety of physicochemical methods elaborated to characterize protein structure and self-organization. These methods include missing electron density in X-ray crystallography maps (16); nuclear magnetic resonance spectroscopy (for recent reviews *see refs. 11 and 17–20* and references therein); circular dichroism spectroscopy in the near-ultraviolet (21) and far-ultraviolet regions (22–25); optical rotatory dispersion spectroscopy (ORD) (22,25); Fourier transform infrared spectroscopy (FTIR) (25); Raman spectroscopy and Raman optical activity (26); fluorescence spectroscopy (27,28); gel-filtration, viscometry, small-angle X-ray scattering, small-angle neutron scattering, sedimentation, and dynamic and static light scattering (27–29); limited proteolysis (30–34); aberrant mobility in sodium dodecyl sulfate-gel electrophoresis (35,36); conformational stability (27,37–40); hydrogen/deuterium exchange (H/D exchange) (28); immunochemical methods (41,42); interaction with molecular chaperones

(27); and electron microscopy or atomic force microscopy. Interested readers can find more detailed description of these approaches in several recent reviews (11,15,18,28).

1.2. Functional Repertoire of Intrinsic Disorder

Ordered proteins have evolved to carry out efficient catalysis and to bind specific ligands. This is complemented by the functional repertoire of IDPs, which are typically involved in regulation, signaling, and control pathways (13,43,44). Using literature searches, Dunker et al. (45) cataloged 90 proteins in which disordered regions were functionally annotated. This group of disordered regions contained 28 specific functions, organized into four functional classes: (1) molecular recognition, (2) molecular assembly, (3) protein modification, and (4) entropic-chain activities (45). Lately, this repertoire was significantly extended applying a novel bioinformatics tool to find functions associated with ID regions (46–48). Using this approach it has been established that out of the 710 Swiss-Prot functional keywords that were associated with at least 20 proteins, 238 were found strongly positively correlated with long ID regions, whereas 302 were strongly negatively correlated (46–48).

1.3. DisProt: A Database of IDPs

Although the first public resource containing disordered protein regions, the ProDDO database, was developed in 2001 (49). This database did not provide information about type of disorder nor the function of disordered regions. Furthermore, it was not curated, being limited to the PDB entries only. These limitations were overcome by currently the most complete database of experimentally characterized disordered proteins, DisProt (50). This database, which can be accessed through <http://www.disprot.org>, provides structural and functional (wherein available) information on experimentally characterized IDPs. As of June 2006, the database contained information on 458 proteins (1096 disordered regions).

1.4. Predicting ID Regions

As already emphasized, bioinformatics played a crucial role in the development of the IDP field. Already at the early stage of the field, simple statistical comparisons of amino acid compositions and sequence complexity indicated that disordered and ordered regions are different to a significant degree. In fact, based on the analysis of 150 ID segments and comparison of these segments with ordered proteins it has been suggested that the amino acids can be grouped into order promoting (C, F, I, L, F, N, V, W, and Y), disorder promoting (A, E, G, K, P, Q, R, and S), and neutral (D, H, M, and T) (10). Several subsequent studies followed up this analysis using increasingly larger data sets (51–54). In addition to the first-order statistics, recent studies also addressed higher-order patterns in amino acid

sequence space and analyzed the space of various physicochemical properties (55), confirming the existence of several biases in IDP sequences. The mentioned sequence biases were exploited to develop a multitude of highly accurate predictors of ID regions, which then were used to estimate the commonness of IDPs in the three kingdoms of life, as well as to elaborate first identifiers of IDP function.

The first predictor of ID regions was reported in 1997 (54). This two-layer feed-forward neural network, which achieved a surprising accuracy of about 70% clearly marked the beginning of a new epoch by showing that (1) there are significant compositional differences between ordered and ID protein regions, (2) the lack of fixed protein 3D structure is predictable from amino acid sequence alone, and (3) ID regions of different lengths (short, medium, and long) may be compositionally different from each other. The predictive model was later extended to the VLXT predictor (51), which is a combination of the VL1 and XT predictors (56). The letters describe the amino acids used for training, where VL stands for Various-characterized Long disordered internal regions and XT stands for X-ray characterized Terminal regions. The VLXT designation is preceded by a descriptive prefix, Predictor of Natural Disordered Regions (PONDR) giving PONDR VLXT.

In 2000, it was noticed that natively unfolded proteins can be separated from ordered proteins by considering their average net charge and hydropathy (25). This observation led to the development of a simple binary classifier, the charge-hydropathy plot (CH-plot) (25), which was based on the analysis of the amino acid composition and instead of predicting ID on a per residue basis, classified entire protein as compact or natively unfolded. Another binary classifier is the cumulative distribution functions (CDF) analysis of disorder scores, which separates ordered and disordered sequences based on the per-residue disorder score retrieved by PONDR VLXT, and the optimal boundary (57,58). This method summarizes the per-residue predictions by plotting PONDR scores against their cumulative frequency, which allows ordered and disordered proteins to be distinguished based on the distribution of prediction scores.

Later, more sophisticated methods based on various statistical and machine-learning techniques (including bagging and boosting [59] and linear regression model for the prediction of long disordered regions [60]) emerged, culminating in the inclusion of the disorder prediction as a separate category in the Critical Assessment of (protein) Structure Prediction (CASP) experiments (61,62). **Table 1** presents the information related to those ID predictors that are scientifically novel and/or published. These predictors are briefly outlined as follows:

1. DISOPRED (63) is a neural network classifier trained on the position-specific scoring matrices and combined disorder prediction with the predictor of secondary structure (64).
2. PONDR VL3 is an ensemble of feed-forward neural networks that uses evolutionary information and is trained on long disordered regions (65).

Table 1
Summary of the Web Servers Offering Prediction of Intrinsically Disordered Proteins

Server name	URL	Approach	References
VLXT (PONDR)	http://www.pondr.com	Feed-forward neural network with separate N-/C-terminus predictor. Based on amino acid compositions and physicochemical properties	51,54,56
FoldIndex©	http://bip.weizmann.ac.il/fldbin/findex	Charge/hydrophobicity score based on a sliding window	25,74
NORSp	http://rostlab.org/services/NORSp/	Rule-based using a set of several neural networks. Amino acid compositions and sequence profiles used as features	68,69
VL2/VL3	http://www.ist.temple.edu/disprot/predictor.php http://www.pondr.com	Ordinary least-squares linear regression (VL2) and bagged feed-forward neural network(VL3). All models use amino acid compositions and sequence complexity. VL3 series uses sequence profiles	2,60,65
DISOPRED	http://bioinf.cs.ucl.ac.uk/disopred/	Feed-forward neural network (DISOPRED) and linear support vector machine (DISOPRED2) based on sequence profiles	63,70,71
GlobPlot	http://globplot.embl.de/	Autoregressive model based on amino acid propensities for disorder/globularity	66
DisEMBL™	http://dis.embl.de/	Ensemble of feed-forward neural networks	67
IUPred	http://iupred.enzim.hu/findex.html	Linear model based on the estimated energy of pairwise interactions in a window around a residue	72,73

(Continued)

Table 1 (Continued)

Server name	URL	Approach	References
PreLink	http://genomics.eu.org/spip/PreLink	Rule-based. Ratio of multinomial probabilities (for linker and structured regions) combined with the distance to the nearest hydrophobic cluster	76
RONN	http://www.w.strubiox.ac.uk/RONN	Feed-forward neural network in the space of distances to a set of prototype sequences of known fold state	75
DISpro	http://www.igb.uci.edu/servers/psss.html	Recursive neural network based on sequence profiles, predicted secondary structure and relative solvent accessibility	77
VSL	http://www.ist.temple.edu/disprot/predictorVSL2.php	Logistic regression (VSL1) and linear support vector machine (VSL2) based on sequence composition, physicochemical properties, and profiles. Combination of short and long disorder predictors	78,79
DRIP-PRED	http://www.sbc.su.se/~maccallr/disorder/	Kohonen's self-organizing maps based on sequence profiles	–
SPRITZ	http://protein.cribi.unipd.it/spritz/	Nonlinear support vector machine based on multiply aligned sequences. Separate predictors for short and long disorder regions	80

3. GlobPlot is based on derived amino acid propensities for disordered regions (66). DisEMBL server uses a support vector machine (67), trained on three proposed types of disorder: (1) loops/coil, i.e., structured regions missing regular secondary structure of helix and strand, (2) hot-loops, i.e., structured regions other than helix or strand, but having high C_{α} B-factors, and (3) remark465, i.e., regions with missing electron density from PDB.
4. NORS predictor identifies regions with nonregular secondary structure (68,69).
5. DISOPRED2 uses linear support vector machines (70,71).
6. IUPred is based on energy-derived coefficients (72,73).
7. FoldIndex (74) is based on the CH approach developed by Uversky et al. (25) and extended to calculations over a sliding window to achieve residue-based predictions.
8. RONN, a regional-order neural network, classifies residues in the space of distances between an input sequence and a set of carefully selected “prototype” sequences (75).
9. PreLink uses compositional bias and lack of hydrophobic clusters (76).
10. DISpro uses large 1D recursive neural networks trained with a variety of compositional, evolutionary, and derived attributes (77).
11. PONDR VSL incorporates the ideas of training separate models for short- (53) and long disordered regions (65), with subsequent combination of these models through a separately trained model (78,79).
12. SPRITZ uses nonlinear support vector machines for short- and long disorder regions based on multiply aligned sequences (80).

Recently, predictors of intrinsic disorder have been used to find functional regions in IDPs. In fact, short regions of predicted order bounded by extended regions of predicted to be disordered by PONDR VLXT, were shown in several cases to identify binding sites that involved disorder-to-order transitions on complex formation (81). These structures, which contained short regions of proteins bound to their partners, showed that the PONDR-indicated region often formed a helix, on binding to its partner. Many examples of these binding sites are found in the PDB (82). The pattern in the PONDR VLXT curve reveals short regions that undergo disorder-to-order transitions on binding. Additionally, these regions tend to have predictions of helix as well as hydrophobic moments. From such characteristics, a predictor of helix-forming molecular recognition features (α -MoRF) was developed (82).

Finally, it has been reported that amino acid compositions, sequence complexity, hydrophobicity, charge, and other sequence attributes of regions adjacent to phosphorylation sites are very similar to those of IDP regions (83). These observations were utilized in the development of a new web-based tool for the prediction of protein phosphorylation sites, disorder-enhanced phosphorylation predictor (DisPhos or DEPP), the accuracy of which reaches 76% for serine, 81% for threonine, and 83% for tyrosine (83).

1.5. When to Use the ID Predictions

In this section some indicators have been outlined regarding when to use the ID predictions, both for the individual protein analysis and for the large-scale studies.

1. ID predictions are priceless for the analysis of individual proteins. These predictions help to better understand and interpret experimental data (e.g., a monomeric protein predicted to be natively unfolded possesses large hydrodynamic volume. Such an unexpectedly large hydrodynamic dimension, being observed experimentally might be incorrectly interpreted in terms of oligomer formation if the protein was assumed to be globular). Such predictions also help to classify proteins and to understand their functionalities. This derives from the observation that the functional repertoires of ordered proteins and IDPs are extremely different. Therefore, knowing that the protein of interest is ID might help redirect its structural and functional analysis. The disorder predictions aided in structural characterization of the retinal tetraspanin (84), nicotinic acetylcholine receptor (85) Dribble, a member of the conserved Krr1P protein family (86) proapoptotic Bcl-2 homology domain-containing family of proteins (87), transcriptional corepressor CtBP (88), notch-signaling pathway proteins (89,90), and many others.
2. Utilizing bioinformatics tools based on ID phenomenon one might find potential protein–protein and protein–nucleic acid interaction sites (molecular recognition fragments) and identify potential sites of posttranslational modifications. This knowledge can be used to drive subsequent research with the major focus on finding binding partners, analysis of resulting complexes, and searching for small molecules modulating these interactions.
3. The majority of ID predictors are based on rather large training sets, which makes prediction of intrinsic disorder in a given protein fairly certain. An ID prediction also means that the analyzed protein is statistically similar to those used in the training of the ID predictors, thus indicating that a particular protein is not an exception, but a rule.
4. ID predictors are indispensable in estimating the commonness of protein disorder in large data sets. They allow scientifically sound extrapolation of knowledge gained on the basis of a few examples to collections including hundreds or even thousands of proteins. For example, proteins associated with cancer (43) and cardiovascular disease (91) were shown to be enriched in intrinsic disorder. ID was shown to be highly abundant in signaling proteins (43), transcription factors (92), proteins with PEST regions (e.g. regions rich in proline, glutamate, serine and threonine) (93), histones (94), serine/arginine-rich splicing factors (95), partners of 14-3-3 proteins (96), nucleoporins (97), and several other sets of proteins with different functions.

Finally, disorder prediction is crucial for protein crystallization and structural genomics projects. Disordered regions are generally not compatible with the crystallization process. Therefore, close examination of sequences that failed to crystallize may reveal ID regions interspersed with regions of order. Thus, accounting for protein disorder can improve target selection and prioritization for the structural genomics projects.

2. Materials

1. The Swiss-Prot database is described in **ref. 98** and is available from <http://www.expasy.org/sprot/>.
2. The database of experimentally characterized disordered proteins, DisProt, is available from <http://www.disprot.org>. The original version of this database is described in **ref. 50**.
3. PONDRVLXT predictor is described in **ref. 51** and is available from <http://www.pondr.com/>.
4. PONDR VL3-BA is described in **ref. 65** and is available from <http://www.pondr.com/>.
5. PONDR VSL is described in **refs. 78 and 79** and is available from <http://www.pondr.com/>.
6. CH-plot predictor is available from <http://www.pondr.com/>. The basic algorithm of this binary classifier is described in **ref. 25**.
7. CDF analysis is available from <http://www.pondr.com/>. This predictor is described in **refs. 57 and 58**.
8. α -MoRF predictor is described in **ref. 82** and is available from <http://www.pondr.com/>.
9. DisPhos predictor also known as DEPP is described in **ref. 83** and is available from <http://www.pondr.com/>.

3. Methods

The methods outlined next describe the analysis of amino acid sequences using the intrinsic disorder knowledge to gain structural and functional information related to a protein, a protein family, or an entire proteome/database. Although numerous predictors of intrinsic disorder are currently available as web servers (*see* DisProt website, <http://www.disprot.org>, for a complete list of such servers), focus will be on utilization of PONDR tools, as they cover a wide range of potential applications of ID concept for structural and functional analysis of proteins. Obviously, this analysis could have been carried out with other ID predictors described earlier.

3.1. Analysis of Protein Amino Acid Composition

It has been already pointed out that a specific feature of a probable ID region is the amino acid compositional bias characterized by a low content of so-called order-promoting residues such as C, V, L, I, M, F, Y, and W and a high content of so-called disorder-promoting residues, including Q, S, P, E, K, G, and A (**10,51,60**). Therefore, the analysis of the amino acid composition biases can provide useful information related to the nature of a given protein. The fractional difference in amino acid composition between a given protein (or a given protein data set) and the set of reference globular proteins is based on the recently elaborated approach (**10**) and provides a perfect visualization tool for

elucidating compositional biases. Here, the fractional difference is calculated as $[f(r) - f_{\text{globular}}(r)]/f_{\text{globular}}(r)$, where $r \in \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}$, $f(r)$ is the frequency of residue r in a given protein set and $f_{\text{globular}}(r)$ is the frequency of residue r in the reference set of globular proteins, and plotted for each amino acid. Negative bars in such a plot correspond to amino acids that are depleted in a given protein in comparison with the set of globular proteins, whereas positive bars reflect the relative increase in the particular amino acid content. Step-by-step design of the fractional difference plot is described next.

3.1.1. Retrieving Sequence Information From the Swiss-Prot Database

Start the Swiss-Prot database by typing <http://www.expasy.org/sprot/> in the Internet browser. Use the following steps to download sequence information in FASTA format.

1. In the window *Search* (located at the top of the front page), choose *Swiss-Prot/TrEMBL* from the pull-down menu. Type the protein name in the neighboring window and click *Go*. Alternatively, click *Full text search* in the UniProt Knowledgebase link located in the *Access to the UniProt Knowledgebase* section of the front page. Type the protein name in the *Enter search terms* window and click *Submit*.
2. On a *search in UniProt Knowledgebase (Swiss-Prot and TrEMBL)* page choose a protein of interest from the list of hits and click the corresponding link.
3. Go to the bottom of the *UniProtKB/Swiss-Prot* entry page and click *FASTA format* link located at the bottom- right corner of the *Sequence Information* section of the page.
4. Copy content of the page, which includes a descriptive header related to the protein and a protein sequence. Keep this information as it will be used in the subsequent analysis. This can be done in Notepad or Microsoft Word. A separate document for each protein is recommended in which all the results of different analyses will be stored.

3.1.2. Applying Proteomic Tools to Obtain Amino Acid Composition

1. Direct approach (if you started with Swiss-Prot database).
 - a. Go to the bottom of the *UniProtKB/Swiss-Prot* entry page and click the *ProtParam* link in *Sequence analysis tools* section.
 - b. *On the ProtParam*: selection of endpoints on the sequence page, click *Submit* if you are going to analyze entire sequence from the previous page. Otherwise, enter the desired endpoints of the sequence in windows provided for *N-* and *C-terminal* points, then hit *Submit*.
 - c. Copy a section of the *ProtParam* page describing *amino acid composition*. Keep this information as it will be used in the subsequent analysis. These are $f(r)$ values for the protein.
2. Alternative approach (if the sequence was retrieved from another source):
 - a. On the Swiss-Prot home page, hit the *Proteomics tools* link located in the top-right corner.

- b. Choose *primary structure analysis* among the several links at the top of the *ExpASY Proteomics tools* page.
- c. Click *ProtParam* link.
- d. Enter a Swiss-Prot/TrEMBL accession number in the space provided or one's own sequence in the box and click *Compute parameters*.
- e. *Copy a section of the ProtParam*: user-provided sequence page describing *amino acid composition*. Keep this information, as it will be used in the subsequent analysis. These are $f(r)$ values; i.e., the frequencies of residue r in the protein.

3.1.3. Compositional Profiling

Table 2 lists averaged frequencies of different residues in a reference set of globular proteins, $f_{\text{globular}}(r)$, and those in a set of experimentally validated IDPs (458 proteins, 1096 disordered regions) from the DisProt database (**50**), $f_{\text{IDP}}(r)$.

1. Rearrange the data for the protein by taking into account that the order of residues you retrieved from the Swiss-Prot is alphabetical (for the three-letter code): Ala(A), Arg(R), Asn(N), Asp(D), Cys(C), Gln(Q), Glu(E), Gly(G), His(H), Ile(I), Leu(L), Lys(K), Met(M), Phe(F), Pro(P), Ser(S), Thr(T), Trp(W), Tyr(Y), and Val(V), whereas it is suggested to list residues according to their disorder propensity, from the least to the most disorder-promoting C, W, Y, I, F, V, L, H, T, N, A, G, D, M, K, R, S, Q, P, and E.
2. Use $f_{\text{globular}}(r)$ values from **Table 2** and $f(r)$ values from the **Subheading 3.1.2**. to calculate the relative frequencies of amino acid residues in the protein as $[f(r) - f_{\text{globular}}(r)]/f_{\text{globular}}(r)$, where $r \in \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}$. This can be done using Excel (Microsoft Corporation, Redmond, WA), SigmaPlot (SYSTAT Software, Inc., San Jose, CA), Origin (OriginLab Corporation, Northampton, MA), or any other graphical software.
3. Use $f_{\text{globular}}(r)$ and $f_{\text{IDP}}(r)$ values from **Table 2** to calculate the relative frequencies of amino acid residues in a set of IDPs as $[f_{\text{IDP}}(r) - f_{\text{globular}}(r)]/f_{\text{globular}}(r)$. This also can be done using Excel, SigmaPlot for Windows, Origin, or any other graphical software.
4. Create a vertical bar chart by plotting the calculated $[f(r) - f_{\text{globular}}(r)]/f_{\text{globular}}(r)$ and $[f_{\text{IDP}}(r) - f_{\text{globular}}(r)]/f_{\text{globular}}(r)$ values for each amino acid residue. For better visual representation, residues should be ranged as follows: C, W, Y, I, F, V, L, H, T, N, A, G, D, M, K, R, S, Q, P, and E; i.e., from the most order-promoting at the left to the most disorder-promoting at the right (*see Fig. 1*).
5. Compare the compositional profiling plot for the protein with that of “averaged” IDP.

Figure 1 illustrates this approach by representing the relative amino acid compositions of the N-terminal (transactivation) domain of the human progesterone receptor (residues 1-566, Swiss-Prot accession no. P06401), protein arginine N-methyltransferase 1 (Swiss-Prot accession no. Q99873), and a set of ID regions available in the DisProt database (**50**). By these computations,

Table 2
Averaged Frequencies of Different Residues (r) in a Reference Set of Globular Proteins, $f_{globular}(r)$, and Those in a Set of Experimentally Validated IDPs (458 Proteins, 1096 Disordered Regions) From the DisProt Database (50), $f_{IDP}(r)$

r	$f_{globular}(r)$	Rmsd	$f_{IDP}(r)$	Rmsd	R	$f_{globular}(r)$	rmsd	$f_{IDP}(r)$	rmsd
W	1.50	0.01	0.78	0.07	T	6.02	0.03	5.33	0.17
C	2.27	0.03	1.04	0.12	R	4.63	0.03	5.64	0.26
F	3.95	0.02	2.45	0.12	G	7.69	0.03	8.43	0.32
I	5.35	0.03	3.28	0.15	Q	3.77	0.02	5.27	0.24
Y	3.70	0.02	2.13	0.14	S	6.29	0.03	8.70	0.28
V	6.88	0.03	4.75	0.17	N	4.53	0.02	4.05	0.19
L	8.34	0.04	6.18	0.20	P	4.80	0.04	6.84	0.28
H	2.33	0.02	2.04	0.11	D	5.43	0.02	6.11	0.23
M	1.92	0.01	2.28	0.12	E	6.09	0.03	8.87	0.32
A	7.98	0.04	8.30	0.30	K	6.23	0.03	7.48	0.30

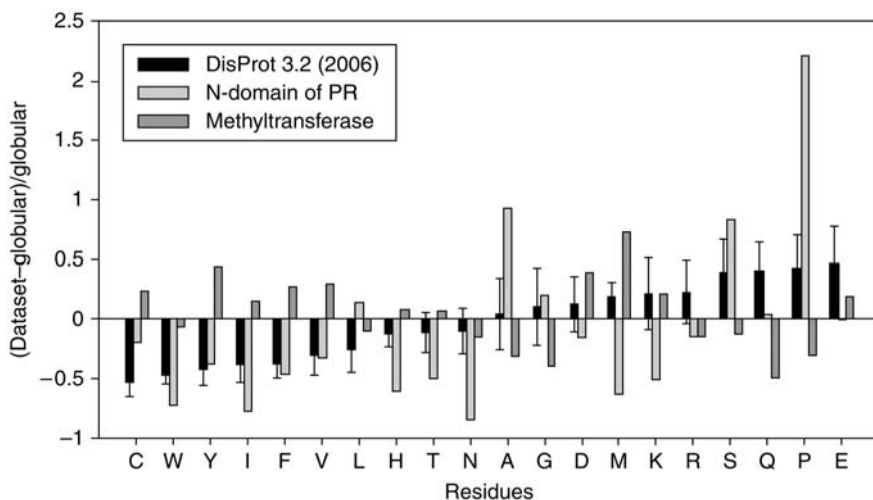


Fig. 1. Amino acid composition, relative to the set of globular proteins globular-3D, of an illustrative IDP, N-terminal (transactivation) domain of the human progesterone receptor (residues 1-566, Swiss-Prot accession no. P06401) (light gray bars); an illustrative ordered protein, protein arginine N-methyltransferase 1 (Swiss-Prot accession no. Q99873) (dark gray bars), and a set of ID regions available in the DisProt 3.2 database (454 proteins, black bars). The arrangement of the amino acids is by peak height for the DisProt 3.2 release. Confidence intervals were estimated using per-protein bootstrapping with 10,000 iterations.

arginine N-methyltransferase 1 is clearly ordered, whereas the transactivation domain is clearly disordered.

3.2. Analyzing Disorder Propensity by PONDR Tools

3.2.1. Entering Information to the PONDR Site and Retrieving Results of ID Prediction

1. Go to the official PONDR site by typing <http://www.pondr.com/> in the Internet browser. You have to be registered to use the bioinformatics tools available at this site. If you are not registered as yet, click *Create a new User Account* link and follow simple instructions there. You will be provided with a username and password. If you are a registered user of PONDR, then click *Log in to a User Account link*, type the username and password in the corresponding windows, and hit *OK*. This will bring you to the PONDR working page.
2. While on the PONDR working page, select boxes corresponding to the desired *Predictors* (VLXT, VL3-BA, VSL1, CDF, and CH). When CH box is marked, two new boxes (*From:* and *To:*) will appear. Leave both empty. Put *Protein name* in the space provided (optional). Enter *NCBI Accession Code* or *Protein Sequence* (FASTA format or sequence only) in the corresponding boxes. Scroll down the page and check

the box *Raw Output* at the *Output Options* section. Clicking *Submit Query* will bring you to the PONDR results page.

3. It is recommended that you keep the content of the entire PONDR results page. Figures can be used as illustrations. *STATISTICS* section provides useful information on the number of residues predicted to be disordered, overall percent of disordered residues, number of disordered regions, the length of the longest disordered region, and the average prediction score. You will find herein a list of regions predicted to be disordered. Raw output values can be used to plot the results for several proteins on one graph.

3.2.2. Understanding the Results of the PONDR Analyses

1. *PONDR scores*. The PONDR results page starts with the plot providing the distribution of PONDR scores over the amino acid sequence. There will be two color lines, blue and red, corresponding to the results for the VLXT and VL3-BA predictions, respectively. Note, when using PONDR VSL1, the results will be emailed. Scores above the threshold of 0.5 correspond to the regions predicted to be disordered. Long disordered regions (with more than 30 consecutive residues predicted to be disordered) are indicated as thick black lines. **Figure 2A1,A2** represent illustrative PONDR score plot for the ID transactivation domain of human progesterone receptor (residues 1-566, Swiss-Prot accession no. P06401) (**Fig. 2A1**) and an ordered protein, protein arginine N-methyltransferase 1 (Swiss-Prot accession no. Q99873) (**Fig. 2A2**). VSL1 curves are added for clarity. The vast majorities of all three curves in **Fig. 2A2** are above the threshold, reflecting the fact that the transactivation domain is highly disordered. Contrarily, the majority of curves for methyltransferase are below the threshold, confirming that this protein is highly ordered. Raw data of these analyses are at the end of the page in the *PREDICTOR VALUES* section.
2. *CDF analysis*. Second plot at the PONDR data page represents the results of CDF analysis. An illustrative CDF curve is shown in **Fig. 2B**. Remember that CDF analysis summarizes the per-residue disorder predictions by plotting PONDR scores against their cumulative frequency, which allows ordered and disordered proteins to be distinguished based on the distribution of prediction scores (57,58). In this case, order-disorder classification is based on whether a CDF curve is above or below a majority of boundary points: if curve is located below the majority of the boundary points (as shown in **Fig. 2B**), then entire protein is predicted to be mostly disordered. However, if the CDF curve is above the boundary, then the analyzed protein is mostly ordered (see **Fig. 2B**). Raw data to reproduce this plot (results for the protein and boundary) are in the *CDF OUTPUT* section.
3. *CH-plot analysis*. The last figure at the PONDR results page shows the CH-plot (25). As aforementioned, compact and natively unfolded proteins plotted in CH space can be separated to a significant degree by a linear boundary, with proteins located above the indicated boundary line being unfolded (red circles) and with proteins below the boundary line being compact (blue squares) (**Fig. 2C**). The protein being tested is marked as a large green square. If this square is above the boundary, then the protein is natively unfolded. If it is below the boundary (as shown

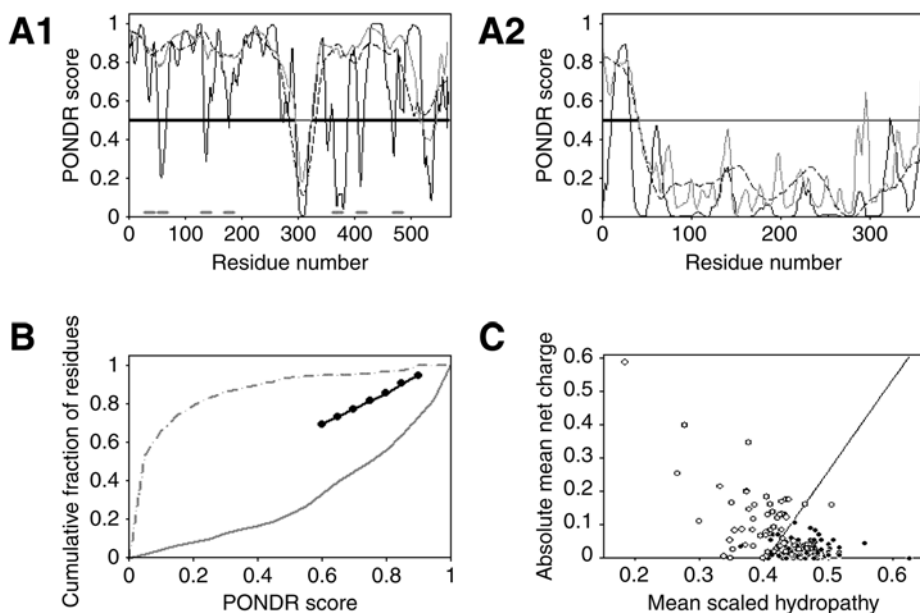


Fig. 2. Illustrative outputs of PONDR algorithms for an illustrative IDP, N-terminal (transactivation) domain of the human progesterone receptor (residues 1-566, Swiss-Prot accession no. P06401) (**Fig. 2A1,B,C**) and an illustrative ordered protein, protein arginine N-methyltransferase 1 (Swiss-Prot accession no. Q99873) (**Fig. 2A2,B,C**). Results of the protein analysis by PONDR VLXT (black solid curves), VL-3B (black dashed), and VSL1 (gray curves) are shown in **Fig. 2A1,A2**. CDF curves for the transactivation domain and methyltransferase are presented in **Fig. 2B** as solid and dashed lines, respectively. **Figure 2C** illustrates corresponding CH-plots, wherein the data for the transactivation domain and methyltransferase are shown as open-crossed square and triangle, respectively. Results of α -MoRF prediction for the transactivation domain of the human progesterone receptor are shown as gray horizontal bars in **Fig. 2A1**. Seven potential α -MoRFs (fragments 27–44, 51–68, 128–145, 168–185, 360–377, 403–420, and 468–485) were identified. *Note*: on your computer screen, results of PONDR and α -MoRF predictions will be present in color: PONDR VLXT will be shown in red, VL-3B in blue, and VSL1 in magenta curves, whereas the results of α -MoRF analysis will be shown as magenta horizontal bars. In CH-plot, data for ordered and natively unfolded proteins are shown as blue squares and red circles, respectively.

in **Fig. 2C**), then the protein is compact. Raw data to build this plot (results for the protein, boundary as well as coordinates of sets of natively unfolded and ordered proteins) are in the *CHARGE-HYDROPATHY OUTPUT* section.

4. *Interpretation of PONDR data* is rather straightforward. As pointed previously, high PONDR scores (more than 0.5) for all three predictors (VLXT, VL3-BA, and VSL1) are characteristic of regions with high propensity to be disordered. Some

peculiarities of the VLXT curve might correlate with protein functionality (see **Subheading 3.3.1**). VL3-BA usually provides very smooth output, as it was trained on long regions of disorder and its raw predictions are averaged over an output window of length 31 to obtain the final prediction for a given position (65). VL3-BA is useful for the accurate prediction of long disordered regions. VSL1 is the most accurate predictor of intrinsic disorder at least in the PONDR series. Its training set is 1335 nonredundant protein sequences, containing 230 long disordered regions with 25,958 residues, 983 short disordered regions with 9632 residues, and 354,169 ordered residues (78,79).

5. *Interpretation of CDF and CH-plot analyses* is straightforward too. It has been pointed out that sometimes these two analyses provide seemingly contradictory data, with CDF analysis predicting a much higher frequency of disorder in sequence databases than CH-plot discrimination (58). The reasons for this discrepancy are outlined in **Subheading 4**. (see **Note 1**). Differences in predictions by these two classifiers were suggested to be physically interpretable in terms of the protein trinity (14) or protein-quartet models (15). Proteins predicted to be disordered by both CH-plot and CDF (i.e., polypeptide chains with high net charge and low hydrophobicity) are likely to be in the extended disorder class. Proteins predicted to be disordered by CDF, but predicted to be ordered by CH-plot, should have properties consistent with a dynamic, collapsed chain and are likely to be in the collapsed disorder class (i.e., molten globules). This supposition needs to be further tested by additional experiments. Rarely, proteins are predicted to be disordered by CH-plot, but ordered by the CDF analysis. This may represent structured proteins with an unusually high net charge; such proteins are likely to exhibit slat-sensitive structures. Finally, proteins predicted to be ordered by both algorithms are of course likely to be in the well-structured class (58). In the application to the illustrative examples of **Fig. 2**, this means that the transactivation domain of human progesterone receptor is most likely a native molten globule, whereas protein arginine N-methyltransferase 1 is likely to be ordered.

3.3. Intrinsic Disorder-Based Functional Analyses

3.3.1. Predicting the Molecular Recognition Fragments, α -MoRFs

The use of disorder predictor to find potential protein-binding sites is based on the observation that the sharp-order dips in otherwise predicted to be disordered regions, could indicate short loosely structured binding regions that undergo disorder-to-order transitions on interaction with the specific binding partner (81). Based on this presumption and the fact that such regions tend to have high α -helical propensities and high hydrophobic moments, a predictor of helix-forming α -MoRF was developed (82). Disorder-to-order transition brings a large decrease in conformational entropy, which is thought to uncouple specificity from binding strength, making highly specific interactions easily reversible. This process is illustrated in **Fig. 3**. The α -MoRF predictor can be accessed at

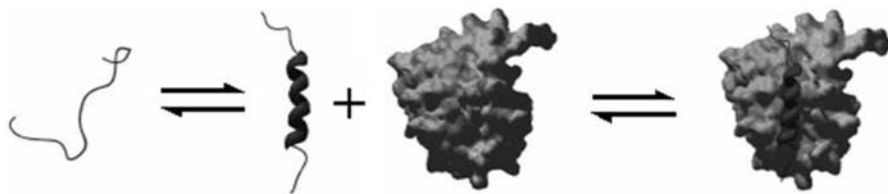


Fig. 3. Illustration of disorder-to-order transition on binding. This example shows the binding of a disordered region of Bad (ribbon) binding to Bcl-XL (globular). Modified from Oldfield et al. (82).

the official PONDR site (<http://www.pondr.com/>) by special request. A typical output of this predictor is shown in **Fig. 2A** as magenta horizontal bars. Notice that the predicted α -MoRFs are located within the distinctive downward spike in the PONDR VLXT curve.

3.3.2. Predicting Potential Phosphorylation Sites

It has been shown that intrinsic disorder prediction might help increase the prediction accuracy of several protein posttranslational modification sites, including protein phosphorylation (83) and methylation (99). For example, DEPP (or DisPhos) uses disorder information to improve the discrimination between phosphorylation and nonphosphorylation sites. The retrieved prediction score approximates the probability that the residue is phosphorylated. Only residues with a prediction score more than 0.5 (which) are considered to be phosphorylated. The step-by-step protocol of DEPP analysis is presented next.

1. Go to the PONDR working page and click the *DEPP Prediction* button. This will bring you to the DEPP working page. While on this page, type *Protein name* in the space provided (optional) and enter *NCBI Accession Code* or *Protein Sequence* (FASTA format or sequence only) in the corresponding boxes. Scroll down the page and check the box *Raw Output* at the *Output Options* section. By clicking *Submit Query* button you will be forwarded to the DEPP results page.
2. The top of DEPP results page represents the plot providing the distribution of DEPP scores over the amino acid sequence. You will have three types of symbols corresponding to the Thr (green triangles), Ser (blue squares), and Tyr residues (red circles) predicted to be phosphorylated. Only residues possessing DEPP scores more than 0.5 are shown. **Fig. 4** represents an illustrative DEPP plot for the transactivation domain of human progesterone receptor (residues 1-566, Swiss-Prot accession no. P06401).
3. Raw data related to this analysis are at the end of the page in the *PREDICTOR VALUES* section. The *DEPP NNP STATISTICS* section provides useful information on the number of phosphorylated serines, threonines, and tyrosines, together with

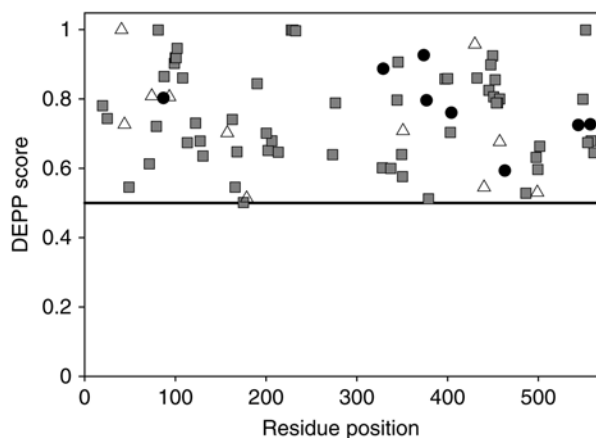


Fig. 4. Prediction of phosphorylation sites in the transactivation domain of the human progesterone receptor (residues 1-566, Swiss-Prot accession no. P06401) by DEPP. The DEPP plot provides the distribution of phosphorylation probability over the amino acid sequence. Symbols corresponding to the Thr (open triangles), Ser (gray squares), and Tyr residues (black circles) predicted to be phosphorylated. Only residues possessing DEPP scores more than 0.5 are shown. *Note:* on your computer screen, results of prediction will be present in color: Thr, Ser, and Tyr residues predicted to be phosphorylated will be shown by green triangles, blue squares, and red circles, respectively.

the total number of these residues in a given protein and the relative phosphorylation efficiency. Once again, it is recommended that one keeps the content of the entire DEPP results page for future use.

4. Notes

1. The difference in the ID prediction by CDF analysis and CH-plot likely results from the fact that the CH-plot is a linear classifier that takes into account only two parameters of the particular sequence—charge and hydrophobicity (25), whereas the CDF analysis is dependent on the output of the PONDR VL-XT predictor, a nonlinear neural network classifier, which was trained to distinguish order and disorder based on a significantly larger feature space that explicitly includes net charge and hydropathy (57,58). Therefore, CH feature space can be considered as a subset of PONDR VL-XT feature space. By definition, CH-plot analysis is predisposed to discriminate proteins with substantial amounts of extended disorder (random coils and premolten globules) from proteins with globular conformations (molten globule-like and rigid well-structured proteins). On the other hand, PONDR-based CDF analysis may discriminate all types of disordered conformations, including molten globules, premolten globules, and coils from ordered proteins (58).

Acknowledgments

The Indiana Genomics Initiative, funded in part by the Lilly Endowment, and National Institute of Health Grant no. 1 R01 LM007688-0A1 provided support for P.R., V.N.U, Z.O, and A.K.D. This work received additional support from the Programs of the Russian Academy of Sciences for the “Molecular and cellular biology” and “Fundamental science for medicine” especially for V.N.U. L.M.I. was supported by National Science Foundation (NSF) grant no. MCB0444818.

References

1. Fischer, E. (1894) Einfluss der configuration auf die wirkung der enzyme. *Ber. Dtsch. Chem. Ges.* **27**, 2985–2993.
2. Obradovic, Z., Peng, K., Vucetic, S., Radivojac, P., Brown, C. J., and Dunker, A. K. (2003) Predicting intrinsic disorder from amino acid sequence. *Proteins* **53**, 566–572.
3. Linderstrom-Lang, K. U. and Schellman, J. A. (1959) Protein structure and enzyme activity, in *The Enzymes*, (Boyer, P. D., Lardy, H., and Myrback, K., eds.), Academic Press, New York, pp. 443–510.
4. Pullen, R. A., Jenkins, J. A., Tickle, I. J., Wood, S. P., and Blundell, T. L. (1975) The relation of polypeptide hormone structure and flexibility to receptor binding: the relevance of X-ray studies on insulins, glucagon and human placental lactogen. *Mol. Cell Biochem.* **8**, 5–20.
5. Cary, P. D., Moss, T., and Bradbury, E. M. (1978) High-resolution proton-magnetic-resonance studies of chromatin core particles. *Eur. J. Biochem.* **89**, 475–482.
6. Holt, C. and Sawyer, L. (1993) Caseins as rheomorphic proteins: interpretation of primary and secondary structures of the α s1-, β -, and κ -caseins. *J. Chem. Soc. Faraday Trans.* **89**, 2683–2692.
7. Schweers, O., Schoenbrunn-Hanebeck, E., Marx, A., and Mandelkow, E. (1994) Structural studies of tau protein and alzheimer paired helical filaments show no evidence for β -structure. *J. Biol. Chem.* **269**, 24,290–24,297.
8. Weinreb, P. H., Zhen, W., Poon, A. W., Conway, K. A., and Lansbury, P. T., Jr. (1996) NACP, a protein implicated in Alzheimer’s disease and learning, is natively unfolded. *Biochemistry* **35**, 13,709–13,715.
9. Wright, P. E. and Dyson, H. J. (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.* **293**, 321–331.
10. Dunker, A. K., Lawson, J. D., Brown, C. J., et al. (2001) Intrinsically disordered protein. *J. Mol. Graph. Model* **19**, 26–59.
11. Daughdrill, G. W., Pielak, G. J., Uversky, V. N., Cortese, M. S., and Dunker, A. K. (2005) Natively disordered protein, in *Protein Folding Handbook*, (Buchner, J. and Kiefhaber, T. eds.), Wiley-VCH: Verlag GmbH & Co., KGaA, Weinheim, pp. 271–353.
12. Uversky, V. N. (2003) A protein-chameleon: conformational plasticity of alpha-synuclein, a disordered protein involved in neurodegenerative disorders. *J. Biomol. Struct. Dyn.* **21**, 211–234.

13. Uversky, V. N., Oldfield, C. J., and Dunker, A. K. (2005) Showing your ID: intrinsic disorder as an ID for recognition, regulation, and cell signaling. *J. Mol. Recognit.* **18**, 343–384.
14. Dunker, A. K. and Obradovic, Z. (2001) The protein trinity-linking function and disorder. *Nat. Biotechnol.* **19**, 805, 806.
15. Uversky, V. N. (2002) Natively unfolded proteins: a point where biology waits for physics. *Protein Sci.* **11**, 739–756.
16. Ringe, D. and Petsko, G. A. (1986) Study of protein dynamics by X-ray diffraction. *Methods Enzymol.* **131**, 389–433.
17. Dyson, H. J. and Wright, P. E. (2002) Insights into the structure and dynamics of unfolded proteins from nuclear magnetic resonance. *Adv. Protein Chem.* **62**, 311–340.
18. Bracken, C., Iakoucheva, L. M., Romero, P. R., and Dunker, A. K. (2004) Combining prediction, computation and experiment for the characterization of protein disorder. *Curr. Opin. Struct. Biol.* **14**, 570–576.
19. Dyson, H. J. and Wright, P. E. (2004) Unfolded proteins and protein folding studied by NMR. *Chem. Rev.* **104**, 3607–3622.
20. Dyson, H. J. and Wright, P. E. (2005) Elucidation of the protein folding landscape by NMR. *Methods Enzymol.* **394**, 299–321.
21. Fasman, G. D. (1996) Circular dichroism and the conformational analysis of biomolecules. Plenum Press, New York.
22. Adler, A. J., Greenfield, N. J., and Fasman, G. D. (1973) Circular dichroism and optical rotatory dispersion of proteins and polypeptides. *Methods Enzymol.* **27**, 675–735.
23. Provencher, S. W. and Glockner, J. (1981) Estimation of globular protein secondary structure from circular dichroism. *Biochemistry* **20**, 33–37.
24. Woody, R. W. (1995) Circular dichroism. *Methods Enzymol.* **246**, 34–71.
25. Uversky, V. N., Gillespie, J. R., and Fink, A. L. (2000) Why are “natively unfolded” proteins unstructured under physiologic conditions? *Proteins* **41**, 415–427.
26. Smyth, E., Syme, C. D., Blanch, E. W., Hecht, L., Vasak, M., and Barron, L. D. (2001) Solution structure of native proteins with irregular folds from Raman optical activity. *Biopolymers* **58**, 138–151.
27. Uversky, V. N. (1999) A multiparametric approach to studies of self-organization of globular proteins. *Biochemistry (Mosc)* **64**, 250–266.
28. Receveur-Brechot, V., Bourhis, J. M., Uversky, V. N., Canard, B., and Longhi, S. (2006) Assessing protein disorder and induced folding. *Proteins* **62**, 24–45.
29. Glatter, O. and Kratky, O. (1982) Small angle X-ray scattering. Academic Press, London.
30. Markus, G. (1965) Protein substrate conformation and proteolysis. *Proc. Natl. Acad. Sci. USA* **54**, 253–258.
31. Mikhalyi, E. (1978) Application of proteolytic enzymes to protein structure studies. CRC Press, Boca Raton.
32. Hubbard, S. J., Eisenmenger, F., and Thornton, J. M. (1994) Modeling studies of the change in conformation required for cleavage of limited proteolytic sites. *Protein Sci.* **3**, 757–768.

33. Fontana, A., de Laureto, P. P., de Filippis, V., Scaramella, E., and Zambonin, M. (1997) Probing the partly folded states of proteins by limited proteolysis. *Fold. Des.* **2**, R17–R26.
34. Fontana, A., de Laureto, P. P., Spolaore, B., Frare, E., Picotti, P., and Zambonin, M. (2004) Probing protein structure by limited proteolysis. *Acta Biochim. Pol.* **51**, 299–321.
35. Iakoucheva, L. M., Kimzey, A. L., Masselon, C. D., Smith, R. D., Dunker, A. K., and Ackerman, E. J. (2001) Aberrant mobility phenomena of the DNA repair protein XPA. *Protein Sci.* **10**, 1353–1362.
36. Tompa, P. (2002) Intrinsically unstructured proteins. *Trends Biochem. Sci.* **27**, 527–533.
37. Privalov, P. L. (1979) Stability of proteins: small globular proteins. *Adv. Protein Chem.* **33**, 167–241.
38. Ptitsyn, O. (1995) Molten globule and protein folding. *Adv. Protein Chem.* **47**, 83–229.
39. Ptitsyn, O. B. and Uversky, V. N. (1994) The molten globule is a third thermodynamical state of protein molecules. *FEBS Lett.* **341**, 15–18.
40. Uversky, V. N. and Ptitsyn, O. B. (1996) All-or-none solvent-induced transitions between native, molten globule and unfolded states in globular proteins. *Fold. Des.* **1**, 117–122.
41. Westhof, E., Altschuh, D., Moras, D., et al. (1984) Correlation between segmental mobility and the location of antigenic determinants in proteins. *Nature* **311**, 123–126.
42. Berzofsky, J. A. (1985) Intrinsic and extrinsic factors in protein antigenic structure. *Science* **229**, 932–940.
43. Iakoucheva, L. M., Brown, C. J., Lawson, J. D., Obradovic, Z., and Dunker, A. K. (2002) Intrinsic disorder in cell-signaling and cancer-associated proteins. *J. Mol. Biol.* **323**, 573–584.
44. Dunker, A. K., Cortese, M. S., Romero, P., Iakoucheva, L. M., and Uversky, V. N. (2005) Flexible nets. The roles of intrinsic disorder in protein interaction networks. *FEBS J.* **272**, 5129–5148.
45. Dunker, A. K., Brown, C. J., Lawson, J. D., Iakoucheva, L. M., and Obradovic, Z. (2002) Intrinsic disorder and protein function. *Biochemistry* **41**, 6573–6582.
46. Xie, H., Vucetic, S., Iakoucheva, L. M., et al. (2007) Functional anthology of intrinsic disorder. I. Biological processes and functions of proteins with long disordered regions. *J. Proteome Res.* **6**, 1882–1898.
47. Vucetic, S., Xie, H., Iakoucheva, L. M., et al. (2007) Functional anthology of intrinsic disorder. 2. Cellular components, domains, technical terms, developmental processes, and coding sequence diversities correlated with long disordered regions. *J. Proteome Res.* **6**, 1899–1916.
48. Xie, H., Vucetic, S., Iakoucheva, L. M., et al. (2007) Functional anthology of intrinsic disorder. 3. Ligands, post-translational modifications and diseases associated with intrinsically disordered proteins. *J. Proteome Res.* **6**, 1917–1932.
49. Sim, K. L., Uchida, T., and Miyano, S. (2001) ProDDO: a database of disordered proteins from the Protein Data Bank (PDB). *Bioinformatics* **17**, 379–380.

50. Vucetic, S., Obradovic, Z., Vacic, V., et al. (2005) DisProt: a database of protein disorder. *Bioinformatics* **21**, 137–140.
51. Romero, P., Obradovic, Z., Li, X., Garner, E. C., Brown, C. J., and Dunker, A. K. (2001) Sequence complexity of disordered protein. *Proteins* **42**, 38–48.
52. Wootton, J. C. (1993) Statistic of local complexity in amino acid sequences and sequence databases. *Comput. Chem.* **17**, 149–163.
53. Radivojac, P., Obradovic, Z., Smith, D. K., et al. (2004) Protein flexibility and intrinsic disorder. *Protein Sci.* **13**, 71–80.
54. Romero, P., Obradovic, Z., Kissinger, C. R., Villafranca, J. E., and Dunker, A. K. (1997) Identifying disordered regions in proteins from amino acid sequences. *IEEE Int. Conf. Neural Netw.* **1**, 90–95.
55. Lise, S. and Jones, D. T. (2005) Sequence patterns associated with disordered regions in proteins. *Proteins* **58**, 144–150.
56. Li, X., Romero, P., Rani, M., Dunker, A. K., and Obradovic, Z. (1999) Predicting protein disorder for N-, C-, and internal regions. *Genome Inform. Ser. Workshop Genome Inform.* **10**, 30–40.
57. Dunker, A. K., Obradovic, Z., Romero, P., Garner, E. C., and Brown, C. J. (2000) Intrinsic protein disorder in complete genomes. *Genome Inform. Ser. Workshop Genome Inform.* **11**, 161–171.
58. Oldfield, C. J., Cheng, Y., Cortese, M. S., Brown, C. J., Uversky, V. N., and Dunker, A. K. (2005) Comparing and combining predictors of mostly disordered proteins. *Biochemistry* **44**, 1989–2000.
59. Vucetic, S., Radivojac, P., Obradovic, Z., Brown, C. J., and Dunker, A. K. (2001) Methods for improving protein disorder prediction, in *International Joint INNS-IEEE Conference on Neural Networks*, Washington, DC, pp. 2718–2723.
60. Vucetic, S., Brown, C. J., Dunker, A. K., and Obradovic, Z. (2003) Flavors of protein disorder. *Proteins* **52**, 573–584.
61. Melamud, E. and Moulton, J. (2003) Evaluation of disorder predictions in CASP5. *Proteins* **53(Suppl 6)**, 561–565.
62. Jin, Y. and Dunbrack, R. L., Jr. (2005) Assessment of disorder predictions in CASP6. *Proteins* **61(Suppl 7)**, 167–175.
63. Jones, D. T. and Ward, J. J. (2003) Prediction of disordered regions in proteins from position specific score matrices. *Proteins* **53**, 573–578.
64. Jones, D. T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292**, 195–202.
65. Peng, K., Vucetic, S., Radivojac, P., Brown, C. J., Dunker, A. K., and Obradovic, Z. (2005) Optimizing long intrinsic disorder predictors with protein evolutionary information. *J. Bioinformatics Comput. Biol.* **3**, 35–60.
66. Linding, R., Russell, R. B., Neduva, V., and Gibson, T. J. (2003) GlobPlot: exploring protein sequences for globularity and disorder. *Nucleic Acids Res.* **31**, 3701–3708.
67. Linding, R., Jensen, L. J., Diella, F., Bork, P., Gibson, T. J., and Russell, R. B. (2003) Protein disorder prediction: implications for structural proteomics. *Structure* **11**, 1453–1459.

68. Liu, J., Tan, H., and Rost, B. (2002) Loopy proteins appear conserved in evolution. *J. Mol. Biol.* **322**, 53–64.
69. Liu, J. and Rost, B. (2003) NORSp: Predictions of long regions without regular secondary structure. *Nucleic Acids Res.* **31**, 3833–3835.
70. Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F., and Jones, D. T. (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.* **337**, 635–645.
71. Ward, J. J., McGuffin, L. J., Bryson, K., Buxton, B. F., and Jones, D. T. (2004) The DISOPRED server for the prediction of protein disorder. *Bioinformatics* **20**, 2138–2139.
72. Dosztanyi, Z., Csizmok, V., Tompa, P., and Simon, I. (2005) The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J. Mol. Biol.* **347**, 827–839.
73. Dosztanyi, Z., Csizmok, V., Tompa, P., and Simon, I. (2005) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* **21**, 3433–3434.
74. Prilusky, J., Felder, C. E., Zeev-Ben-Mordehai, T., et al. (2005) FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics* **21**, 3435–3438.
75. Yang, Z. R., Thomson, R., McNeil, P., and Esnouf, R. M. (2005) RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics* **21**, 3369–3376.
76. Coeysaux, K. and Poupon, A. (2005) Prediction of unfolded segments in a protein sequence based on amino acid composition. *Bioinformatics* **21**, 1891–1900.
77. Cheng, J., Sweredoski, M. J., and Baldi, P. (2005) Accurate prediction of protein disordered regions by mining protein structure data. *Data Mining Knowledge Disc.* **11**, 213–222.
78. Obradovic, Z., Peng, K., Vucetic, S., Radivojac, P., and Dunker, A. K. (2005) Exploiting heterogeneous sequence properties improves prediction of protein disorder. *Proteins* **61(Suppl 7)**, 176–182.
79. Peng, K., Radivojac, P., Vucetic, S., Dunker, A. K., and Obradovic, Z. (2006) Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics* **7**, 208.
80. Vullo, A., Bortolami, O., Pollastri, G., and Tosatto, S. C. (2006) Spritz: a server for the prediction of intrinsically disordered regions in protein sequences using kernel machines. *Nucleic Acids Res.* **34**, W164–W168.
81. Garner, E., Romero, P., Dunker, A. K., Brown, C., and Obradovic, Z. (1999) Predicting binding regions within disordered proteins. *Genome Inform. Ser. Workshop Genome Inform.* **10**, 41–50.
82. Oldfield, C. J., Cheng, Y., Cortese, M. S., Romero, P., Uversky, V. N., and Dunker, A. K. (2005) Coupled folding and binding with alpha-helix-forming molecular recognition elements. *Biochemistry* **44**, 12,454–12,470.
83. Iakoucheva, L. M., Radivojac, P., Brown, C. J., et al. (2004) The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res.* **32**, 1037–1049.

84. Ritter, L. M., Arakawa, T., and Goldberg, A. F. (2005) Predicted and measured disorder in peripherin/rds, a retinal tetraspanin. *Protein Pept. Lett.* **12**, 677–686.
85. Kukhtina, V., Kottwitz, D., Strauss, H., et al. (2005) Intracellular domain of nicotinic acetylcholine receptor: the importance of being unfolded. *J. Neurochem.*
86. Yiu, C. P., Beavil, R. L., and Chan, H. Y. (2006) Biophysical characterisation reveals structural disorder in the nucleolar protein, Dribble. *Biochem. Biophys. Res. Commun.* **343**, 311–318.
87. Hinds, M. G., Smits, C., Fredericks-Short, R., et al. (2007) Bim, Bad and Bmf: intrinsically unstructured BH3-only proteins that undergo a localized conformational change on binding to prosurvival Bcl-2 targets. *Cell Death Differ.* **14**, 128–136.
88. Nardini, M., Svergun, D., Konarev, P. V., et al. (2006) The C-terminal domain of the transcriptional corepressor CtBP is intrinsically unstructured. *Protein Sci.* **15**, 1042–1050.
89. Roy, S., Schnell, S., and Radivojac, P. (2007) Unraveling the nature of the segmentation clock: intrinsic disorder of clock proteins and their interaction map. *Comput. Biol. Chem.* **30**, 241–248.
90. Popovic, M., Coglievina, M., Guarnaccia, C., et al. (2006) Gene synthesis, expression, purification, and characterization of human Jagged-1 intracellular region. *Protein Expr. Purif.* **47**, 398–404.
91. Cheng, Y., Le Gall, T., Oldfield, C. J., Dunker, A. K., and Uversky, V. N. (2006) Abundance of intrinsic disorder in proteins associated with cardiovascular disease. *Biochemistry* **45**, 10,448–10,460.
92. Liu, J., Perumal, N. B., Oldfield, C. J., Su, E. W., Uversky, V. N., and Dunker, A. K. (2006) Intrinsic disorder in transcription factors. *Biochemistry* **45**, 6873–6888.
93. Singh, G. P., Ganapathi, M., Sandhu, K. S., and Dash, D. (2006) Intrinsic unstructuredness and abundance of PEST motifs in eukaryotic proteomes. *Proteins* **62**, 309–315.
94. Hansen, J. C., Lu, X., Ross, E. D., and Woody, R. W. (2006) Intrinsic protein disorder, amino acid composition, and histone terminal domains. *J. Biol. Chem.* **281**, 1853–1856.
95. Haynes, C. and Iakoucheva, L. M. (2006) Serine/arginine-rich splicing factors belong to a class of intrinsically disordered proteins. *Nucleic Acids Res.* **34**, 305–312.
96. Bustos, D. M. and Iglesias, A. A. (2006) Intrinsic disorder is a key characteristic in partners that bind 14-3-3 proteins. *Proteins* **63**, 35–42.
97. Denning, D. P., Patel, S. S., Uversky, V., Fink, A. L., and Rexach, M. (2003) Disorder in the nuclear pore complex: the FG repeat regions of nucleoporins are natively unfolded. *Proc. Natl. Acad. Sci. USA* **100**, 2450–2455.
98. Boeckmann, B., Bairoch, A., Apweiler, R., et al. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31**, 365–370.
99. Daily, K. M., Radivojac, P., and Dunker, A. K. (2005) Intrinsic disorder and protein modifications: building an SVM predictor for methylation, in *IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, CIBCB 2005*, San Diego, California, CA, pp. 475–481.