Published in final edited form as:

Protein Pept Lett. 2008 ; 15(9): 956–963.

TOP-IDP-Scale: A New Amino Acid Scale Measuring Propensity for Intrinsic Disorder

Andrew Campen¹, Ryan M. Williams², Celeste J. Brown³, Jingwei Meng⁴, Vladimir N. Uversky^{4,5,6,*}, and A. Keith Dunker^{4,*}

¹Department of Computer and Information Science, Indiana University-Purdue University Indianapolis, Indianapolis, IN 46202, USA

²School of Electrical Engineering and Computer Science, Washington State University, Pullman, WA 99164, USA

³Initiative for Bioinformatics and Evolutionary Studies, University of Idaho, Moscow, ID 83844, USA

⁴Center for Computational Biology and Bioinformatics, Department of Biochemistry and Molecular Biology, Indiana University School of Medicine, Indianapolis, IN 46202, USA

⁵Institute for Biological Instrumentation, Russian Academy of Sciences, 142290 Pushchino, Moscow Region, Russia

⁶Institute for Intrinsically Disordered Protein Research, Indiana University School of Medicine, Indianapolis, IN 46202, USA

Abstract

Intrinsically disordered proteins carry out various biological functions while lacking ordered secondary and/or tertiary structure. In order to find general intrinsic properties of amino acid residues that are responsible for the absence of ordered structure in intrinsically disordered proteins we surveyed 517 amino acid scales. Each of these scales was taken as an independent attribute for the subsequent analysis. For a given attribute value X , which is averaged over a consecutive string of amino acids, and for a given data set having both ordered and disordered segments, the conditional probabilities $P(s_o | x)$ and $P(s_d | x)$ for order and disorder, respectively, can be determined for all possible values of X . Plots of the conditional probabilities $P(s_o | x)$ and $P(s_d | x)$ versus X give a pair of curves. The area between these two curves divided by the total area of the graph gives the area ratio value (ARV), which is proportional to the degree of separation of the two probability curves and, therefore, provides a measure of the given attribute's power to discriminate between order and disorder. As ARV falls between zero and one, larger ARV corresponds to the better discrimination between order and disorder. Starting from the scale with the highest ARV, we applied a simulated annealing procedure to search for alternative scale values and have managed to increase the ARV by more than 10%. The ranking of the amino acids in this new TOP-IDP scale is as follows (from order promoting to disorder promoting): W, F, Y, I, M, L, V, N, C, T, A, G, R, D, H, Q, K, S, E, P. A web-based server has been created to apply the TOP-IDP scale to predict intrinsically disordered proteins (<http://www.disprot.org/dev/disindex.php>).

*To whom correspondence should be addressed: VNU, Department of Biochemistry and Molecular Biology, Indiana University School of Medicine, 410 W. 10th Street, HS 5000, Indianapolis, IN 46202, USA; Phone: 317-278-6448; Fax: 317-278-9217; E-mail: E-mail: vuversky@iupui.edu; AKD, Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, 410 W. 10th Street, HS 5000, Indianapolis, IN 46202, USA; Phone: 317-278-9650; Fax: 317-278-9217; E-mail: E-mail: kedunker@iupui.edu.

Keywords

intrinsic disorder; amino acid scale; conditional probability

INTRODUCTION

Numerous pieces of experimental data support the view that many biologically active proteins (known as intrinsically disordered proteins, IDPs) remain unstructured, or incompletely structured, under physiological conditions (reviewed in [1-15]). The functions carried out by IDPs are diverse and complement those of ordered protein regions. This hypothesis has been recently supported by a comprehensive analysis of the correlation of nearly all keywords in SwissProt database with over- or under-prediction of intrinsic disorder [16-18]. Out of the 710 Swiss-Prot functional keywords that were associated with at least 20 proteins, 238 were found to be strongly positively correlated with predictions of long intrinsically disordered regions, whereas 302 were strongly negatively correlated with such predictions. The negatively correlated proteins are likely to be structured, and the top-ranking keywords for these proteins all end in “ase,” and thus these keywords describe enzymes the functions of which depend on active sites resulting from folding into globular 3D structures. The disorder-associated function list is rich in keywords describing signaling, regulation, and control. These functions often rely on interactions with multiple partners where high-specificity/low-affinity interactions can result [16-18].

ID can be manifested in a variety of contexts, affecting various levels of protein structure: functional disordered segments can be as short as only a few amino acid residues, or they can occupy rather long loop regions and/or protein ends. Proteins can be partially or even wholly disordered, even large ones. Disordered proteins and regions have been grouped into at least two broad structural classes — compact (molten globule-like) and extended (coil-like and pre-molten globule-like, so called natively unfolded proteins). In general, amino acid sequences encoding for the disordered proteins or regions are significantly different from those characteristic for the ordered proteins [2,4,19-22]. This implies that in addition to the well-known “protein folding code” stating that all the information necessary for a given protein to fold is encoded in its amino acid sequence [23], there is also a “protein non-folding code”, according to which the propensity of a protein to stay intrinsically disordered is encoded in its amino acid sequence too [24]. In fact, IDPs are significantly depleted in bulky hydrophobic (Ile, Leu, and Val) and aromatic amino acid residues (Trp, Tyr, and Phe), which would normally form the hydrophobic core of a folded globular protein, and also possess low content of Cys and Asn residues. These depleted residues, Trp, Tyr, Phe, Ile, Leu, Val, Cys and Asn were proposed to be called order-promoting amino acids. On the other hands, natively unfolded proteins were shown to substantially enriched in polar, disorder-promoting, amino acids: Ala, Arg, Gly, Gln, Ser, Pro, Glu, and Lys [4,15,24,25].

Given the existence and abundance of intrinsically disordered proteins, a focus of many researchers has been to understand the relationship between amino acid sequence (or composition) and protein folding or non-folding and to use this knowledge to develop predictors of intrinsic disorder. The first such predictor was constructed in 1997 by Romero et al. based only on 67 disordered regions (1,340 residues) and a number of ordered regions (16,543 residues) manually extracted from PDB [21]. Based on these data, a two-layer feed-forward neural-network was constructed that achieved a surprising accuracy of about 70%. The predictive model was later extended into the PONDR VLXT predictor [25], a combination of an interior disordered region predictor (VL1) and a separate predictor trained only at protein termini, XT [26]. In 2000, Uversky et al. noticed that proteins disordered over their entire lengths can be separated from ordered proteins by considering their average net charge and

hydropathy (a scale describing the relative hydrophobicity/hydrophilicity of amino acid residues) [2]. In its original form, the charge-hydropathy plot (CH-plot) did not have the sensitivity to predict disordered regions on a per residue basis, but recently the CH- analysis has been modified and extended to identify local ID regions using a sliding window approach [27]. In time, more sophisticated methods based on various statistical and machine learning techniques have emerged. The predictors developed so far have been based on a spectrum of computational approaches relying on amino acid compositions, derived properties (such as secondary structure prediction) or simple physicochemical properties (such as charge) of the local sequence neighborhood. Almost all of the above-mentioned predictors are available as web servers. Links to these servers, when available, can be found in the Disordered Protein Database or DisProt for short (www.disprot.org) [28,29].

The goal of this study was to construct a new amino acid attribute, e.g. a new amino acid scale that discriminates between order and disorder. To this end, 517 amino acid attributes (including a variety of hydrophobicity scales, different measures of side chain bulkiness, polarity, volume, compositional attributes, the frequency of each single amino acid and so on) were analyzed using an approach based on the analysis of the area ratio value (ARV) between conditional probability curves described in detail previously [24,30]. The new scale developed here, TOP-IDP scale, out-performed the other 517 amino acid scales for the discrimination of order and disorder. This scale provides a new ranking for the tendencies of the amino acid residue to promote order or disorder. It was used to create a web-based server for the prediction of intrinsic disorder in proteins. This server is freely available at <http://www.disprot.org/dev/disindex.php>.

MATERIALS AND METHODS

Databases

The creation and evaluation of sequence attributes to discriminate between order and disorder requires a database containing a balanced set of both ordered and disordered residues. Disordered regions identified by NMR, circular dichroism, or protease digestion were found by keyword searches of PubMed (<http://www.ncbi.nlm.nih.gov>). Additionally, starting from a subset of the Protein Data Bank (PDB) called PDB_Select_25, disordered regions in X-ray crystal structures were identified by searching for residues having backbone atoms that are absent from the ATOMS lists in their PDB files. These differently characterized disordered regions were grouped together in into a set called dis_ALL. In total, dis_ALL contains 154 proteins and 92,735 residues. An ordered set of proteins was created using the non-redundant subset of PDB called PDB_Select_25. Proteins without missing atom coordinated were selected and in total, the ordered set contains 290 proteins and 67,548 residues.

Amino acid sequence attributes

The values for various amino properties such as hydrophobicity, polarity, volume, etc., were compiled by database and literature searches. The AAIndex database (<http://www.genome.jp/aaindex/>) provided 494 distinct numerical property scales. In addition, three compositional attributes, and the frequency of each single amino acid were also included in the complete list. Disregarding high correlation between certain scales, all 517 distinct property scales were examined.

Scoring function for order / disorder discrimination

In this study, a method for determining how well one property scale discriminates between order/disorder versus another property scale was needed. For this evaluation, we used the sequence attributes area ratio method, which was described in detail previously [24,30].

The first step in the evaluation method was to balance disordered windows of 21 residues from the disordered database with an equal number of ordered windows of the same size taken from the ordered dataset. Next, given a balanced number of ordered and disordered windows, the attribute values for each window were calculated as averages over the window and the windows were then binned by their attribute values. From the proportion of ordered and disordered windows in each bin, the conditional probability curves of order and disorder, $P(s_o | x)$ and $P(s_d | x)$ respectively, were estimated. The conditional probability curves were then plotted versus the attribute values associated with each bin. Next, the relative degree of separation of the two probability curves was determined by calculating the area between the order and disorder curves and dividing by the area of the graph, giving the area ratio value (ARV). The ARV falls between zero and one with the larger the ARV, the better the discrimination between order and disorder.

Algorithms to optimize a scale for order / disorder discrimination

Due to the large size of the attribute space, we used two different genetic algorithms to efficiently find an optimal scale for order/disorder discrimination. Genetic algorithms are a particular class of evolutionary algorithms that include techniques inspired by evolutionary biology such as inheritance, mutation, selection, and crossover/recombination. For our study, each attribute property scale is considered as an individual. As described above, we used the sequence attribute method as the fitness function meaning individuals having higher ARVs being considered of higher quality than ones with lower ARVs. The basic pseudo-code genetic algorithm looks like this:

```

Choose initial population
Repeat
  Evaluate the fitness of each individual in the population.
  Select a certain portion of the best-ranking individuals to reproduce.
  Breed a new population through crossover/recombination and mutation.
Until terminating condition
  
```

Specifically, both a Canonical Genetic Algorithm (CGA) [31] and a Stochastic Hill Climbing Algorithm (SHCA) [32] were tested. The SHCA began with the initial set of 517 property scales as the initial population. After this population was evaluated and ranked, the top-ranked scale was selected to create the new population. The first weight value in this scale was then manipulated 99 times by adding a small random value to this initial weigh value. The new 100-scale population, including the original top-ranked scale, was again evaluated and a new top-ranked scale selected. Successive iterations manipulated the next weight value, viewed as a circular list of values, with decreasing variability to bring about convergence. The algorithm terminated when the improvement in fitness between iterations failed to change significantly.

The CGA started with population of 100 scales randomly selected from the list of 517. The scales were evaluated and ranked using the same sequence attribute method described above. The reproduction fitness of each scale is determined by $rf_i = f_i / f_{av}$, where f_i is the fitness value of the current individual and f_{av} is the average fitness value for all the members of the population. By using “remainder stochastic sampling” [31], an intermediate population of 100 scales was constructed. The fractional portion of the reproductive fitness value, rf_i , was used to determine the likelihood of the individual being selected for the next population while the integer portion was used to determine the number of times that individual will appear in the

next population if it is selected. Scales are randomly selected based on their likelihood until the intermediate population is full.

Next, pairs of scales in the intermediate population were randomly selected for recombination with a probability of 90%. Crossover/recombination was achieved by swapping randomly selected parts of each parent scale to generate two new scales, each having a part of both parents, which were placed in the next population. The algorithm selected one random point at which to split and recombine. After crossover/recombination, mutation was applied with a probability of 5% for each weight in each scale. Mutation took place by manipulating a given weight by a random amount between 1 and -1. Once these operations were complete, a new iteration of the algorithm began until the fitness of the top-ranked scale failed to improve significantly.

RESULTS AND DISCUSSION

Ranking previously published attributes

Using the sequence attribute method described above, ARVs were calculated for 517 previously published attributes. The ARVs for the top 10 attribute scales are listed in Table 1. The rankings of all 517 scales are available by request. The highest ranked attribute is the normalized flexibility parameters (B-values) for each residue surrounded by two rigid neighbors [33] with an ARV of 0.687. All other ARVs were lower than this, ranging as low as 0.072 for composition of amino acids in extracellular proteins (percent) [34]. Figure (1) illustrates the area ratio comparison graph for the highest value (Figure (1A)). Normalized flexibility parameters (B-values) for each residue surrounded by two rigid neighbors [33]), average value (Figure (1B)). Positive charge [35]), and lowest value (Figure (1C)), composition of amino acids in extracellular proteins, percent [34]). It can be seen that the area between the two curves is greater in Figure (1A) than both Figure (1B) and (1C).

TOP-IDP, an optimized scale for order / disorder discrimination

Our optimization techniques, SHCA and CGA, both resulted in scales that were essentially identical. The values of this newly optimized scale, called TOP-IDP, are compared with the best of the published scales, normalized flexibility parameters (B-values) for each residue surrounded by one rigid neighbor [33], in Table 2. Amino acids were ordered according to the scale values of the TOP-IDP scale and were divided into order-promoting and disorder-promoting categories. Note that both of these scales are over difference ranges, so comparison is difficult. When ranked according to ARV with the original 517 published attributes, TOP-IDP gave highest-ranked scale with an ARV of 0.761 as compared to 0.687 for the scale that previously ranked highest. This represents an 11% improvement in the ability to discriminate between order and disorder. Figure (1D) shows the area ratio graph for the TOP-IDP scale.

Comparing four scales

Interestingly, the only disorder propensity scale developed so far was based on the ability of amino acid residues to form a sufficient number of contacts in a globular state [36]. In this approach, the average number of contacts per residue was calculated for each residue based on a set of ordered proteins, giving rise to the FoldUnfold scale (see Table 2), and then the expected average number of contacts per residue calculated from the amino acid sequence alone (using the average number of contacts for 20 amino acid residues in globular proteins) was used as an indicator of natively unfolded proteins [36]. Another disorder propensity scale can be derived from the relative amino acid composition of IDPs. To this end, the fractional difference in composition between a set of IDPs deposited in DisProt [28,29] and a set of ordered proteins from PDB is calculated for each amino acid residues as $(C_{\text{disorder}} - C_{\text{order}}) / C_{\text{order}}$, where C_{disorder} is the averaged content of a given amino acid in a set of IDPs and C_{order} is the corresponding averaged content in a set of ordered proteins [37]. These data can

be used to derive a disorder propensity scale, DisProt scale, which is based on the statistical difference in the amino acid compositions of ordered proteins and IDPs.

The ranking of the amino acids for TOP-IDP, normalized flexibility parameters (B-values) for each residue surrounded by one rigid neighbor [33], FoldUnfold scale and DisProt scale are compared graphically in Figure (2A). To obtain this data, all the scales were normalized to have the minimal value of zero and the maximal value of 1. The order of the amino acids along the X-axis followings the ranking of TOP-IDP as can be seen from the monotonically increasing values from left to right. If the amino acid were ranked by the B-values, FoldUnfold or DisProt scales, the order of the amino acids would be very different. Figure (2A) shows that for any amino acid residues the unfoldability score assessed by Top-IDP scale is comparable with score assessed by at least one other scale, suggesting that these four scales are correlated. Figure (2B-G) represents the pair-wise comparison of these four scales and shows that they possess significant mutual correlation. Interestingly, the greatest correlation ($r^2 = 0.785$) was observed for the TOP-IDP-FoldUnfold pair. This is rather interesting finding, as four scales compared were derived using absolutely different models - normalized flexibility parameters (B-values) for each residue surrounded by one rigid neighbor [33], expected average number of contacts per residue [36], statistical difference in the amino acid compositions of intrinsically disordered and ordered proteins, and a novel amino acid attribute constructed for the discrimination between order and disorder.

TOP-IDP web-based prediction service

To apply the TOP-IDP scale to prediction of intrinsically disordered protein, a web-based prediction service has been created at <http://www.disprot.org/dev/disindex.php>. The page takes in a protein sequence, in either 1-letter or 3-letter code FASTA format, as input. The average global TOP-IDP value and average window-by-window TOP-IDP values are calculated based on the normalized TOP-IDP scale. Based on maximum-likelihood methods, a prediction cut-off of 0.542 was calculated giving the equation: $I_{Top-IDP} = -(<Top - IDP> - 0.542)$ where $<Top - IDP>$ is the average TOP-IDP value for a protein (or window). Positive prediction values indicate proteins (or windows) that are likely to be ordered, were negative prediction values indicate proteins (or windows) what are likely to be intrinsically disordered. In order to better interpret prediction results, window prediction are plotted sequentially with the prediction value representing the center residue of the given window.

For an additional comparison, the prediction service also implements the improved algorithm of Uversky et al. [2] which used a combination of Kyte/Doolittle hydrophathy and net charge to predict intrinsic disorder [27]. These values are also plotted in the same method as the TOP-IDP value for easy comparison. All data are available for download via either tab-delimited text or comma-separated text (csv) format.

Example of application of TOP-IDP server

Figure (3) represents an illustrative example of the Top-IDP server application. Here, the disorder propensity of the tumor suppressor protein p53 (which is at the center of a large signaling network, regulating expression of genes involved in many cellular processes such as cell cycle progression, apoptosis induction, DNA repair, and response to cellular stress [38] is estimated by Top-IDP and two well-accepted disorder predictors, PONDR® VLXT and VSL1. There are three structural domains in p53: N-terminal translational activation domain, central DNA binding domain, and C-terminal tetramerization and regulatory domain. At the transactivation region, it interacts with TFIID, TFIIH, Mdm2, RPA, CBP/p300 and CSN5/Jab1 [38]. At the C-terminal domain, it interacts with GSK3 β , PARP-1, TAF1, TRRAP, hGcn5, TAF, 14-3-3, and S100B($\beta\beta$). Both N- and C-terminal domains of p53 are involved in numerous protein-protein interactions, some of which involve disorder-to-order transitions. PONDR®

VLXT plot shown in Figure (3B) illustrates such a predisposition for the disorder-to-order transition as sharp dips — or short regions of predicted order — within longer regions of predicted disorder. Recently we have established that this capability of PONDR® VLXT to find the potential protein-protein interaction sites involving disorder-to-order transitions is rather unique among more than 15 predictors analyzed [46]. Intriguingly, TOP-IDP was able to reproduce many of the characteristic features observed in the PONDR® VLXT plot, suggesting that this new predictor has some interesting potential applications.

ACKNOWLEDGEMENTS

This work was supported in part by the grants R01 LM007688-01A1 (to A.K.D and V.N.U.) and GM071714-01A2 (to A.K.D and V.N.U.) from the National Institutes of Health and the Programs of the Russian Academy of Sciences for the “Molecular and cellular biology” and “Fundamental science for medicine” (to V. N. U.). We gratefully acknowledge the support of the IUPUI Signature Centers Initiative.

Biography

Dr. Uversky received broad training, with an MS in Physics, and a PhD and a DSc in Biophysics. He is using molecular biophysics and bioinformatics methods to study protein folding, misfolding and nonfolding. Dr. Uversky published > 250 research papers and reviews and edited several scientific books and book series.

List of abbreviations

ID, intrinsic disorder; IDP, intrinsically disordered protein; ARV, area ratio value; PDB, protein data bank; PONDR, predictor of natural disordered regions; CGA, canonical genetic algorithm; SHCA, stochastic hill climbing algorithm; CH-plot, charge-hydropathy plot.

REFERENCES

- [1]. Wright PE, Dyson HJ. *J Mol Biol* 1999;293:321–331. [PubMed: 10550212]
- [2]. Uversky VN, Gillespie JR, Fink AL. *Proteins* 2000;41:415–427. [PubMed: 11025552]
- [3]. Dunker AK, Obradovic Z. *Nat Biotechnol* 2001;19:805–806. [PubMed: 11533628]
- [4]. Dunker AK, Lawson JD, Brown CJ, Williams RM, Romero P, Oh JS, Oldfield CJ, Campen AM, Ratliff CM, Hipps KW, Ausio J, Nissen MS, Reeves R, Kang C, Kissinger CR, Bailey RW, Griswold MD, Chiu W, Garner EC, Obradovic Z. *J Mol Graph Model* 2001;19:26–59. [PubMed: 11381529]
- [5]. Uversky VN. *Protein Sci* 2002;11:739–756. [PubMed: 11910019]
- [6]. Uversky VN. *Eur J Biochem* 2002;269:2–12. [PubMed: 11784292]
- [7]. Uversky VN. *Cell Mol Life Sci* 2003;60:1852–1871. [PubMed: 14523548]
- [8]. Dunker AK, Brown CJ, Obradovic Z. *Adv Protein Chem* 2002;62:25–49. [PubMed: 12418100]
- [9]. Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradovic Z. *Biochemistry* 2002;41:6573–6582. [PubMed: 12022860]
- [10]. Iakoucheva LM, Brown CJ, Lawson JD, Obradovic Z, Dunker AK. *J Mol Biol* 2002;323:573–584. [PubMed: 12381310]
- [11]. Tompa P. *Trends Biochem Sci* 2002;27:527–533. [PubMed: 12368089]
- [12]. Daughdrill, GW.; Pielak, GJ.; Uversky, VN.; Cortese, MS.; Dunker, AK. *Handbook of Protein Folding*. Buchner, J.; Kiefhaber, T., editors. Wiley-VCH, Verlag GmbH & Co. KGaA; Weinheim, Germany: 2005. p. 271-353.
- [13]. Uversky VN, Oldfield CJ, Dunker AK. *J Mol Recognit* 2005;18:343–384. [PubMed: 16094605]
- [14]. Dunker AK, Cortese MS, Romero P, Iakoucheva LM, Uversky VN. *Febs J* 2005;272:5129–48. [PubMed: 16218947]
- [15]. Radivojac P, Iakoucheva LM, Oldfield CJ, Obradovic Z, Uversky VN, Dunker AK. *Biophys J* 2007;92:1439–56. [PubMed: 17158572]
- [16]. Vucetic S, Xie H, Iakoucheva LM, Oldfield CJ, Dunker AK, Obradovic Z, Uversky VN. *J Proteome Res* 2007;6:1899–916. [PubMed: 17391015]
- [17]. Xie H, Vucetic S, Iakoucheva LM, Oldfield CJ, Dunker AK, Obradovic Z, Uversky VN. *J Proteome Res* 2007;6:1917–32. [PubMed: 17391016]
- [18]. Xie H, Vucetic S, Iakoucheva LM, Oldfield CJ, Dunker AK, Uversky VN, Obradovic Z. *J Proteome Res* 2007;6:1882–98. [PubMed: 17391014]
- [19]. Dunker AK, Garner E, Guillot S, Romero P, Albrecht K, Hart J, Obradovic Z, Kissinger C, Villafranca JE. *Pac Symp Biocomput* 1998:473–484. [PubMed: 9697205]
- [20]. Romero P, Obradovic Z, Dunker AK. *Genome Informatics* 1997;8:110–124. [PubMed: 11072311]
- [21]. Romero, P.; Obradovic, Z.; Kissinger, C.; Villafranca, JE.; Dunker, AK. 1997 *Proceedings of International Conference on Neural Networks*; 1997. p. 90-95.
- [22]. Romero P, Obradovic Z, Kissinger CR, Villafranca JE, Garner E, Guillot S, Dunker AK. *Pac Symp Biocomput* 1998:437–448. [PubMed: 9697202]
- [23]. Creighton TE. *Science* 1988;240:267, 344. [PubMed: 3353718]
- [24]. Williams RM, Obradovic Z, Mathura V, Braun W, Garner EC, Young J, Takayama S, Brown CJ, Dunker AK. *Pac Symp Biocomput* 2001:89–100. [PubMed: 11262981]
- [25]. Romero P, Obradovic Z, Li X, Garner EC, Brown CJ, Dunker AK. *Proteins* 2001;42:38–48. [PubMed: 11093259]
- [26]. Li X, Romero P, Rani M, Dunker AK, Obradovic Z. *Genome Inform Ser Workshop Genome Inform* 1999;10:30–40.

- [27]. Priluski J, Felder CE, Zeev-Ben-Mordehai T, Rydberg EH, Man O, Beckman JS, Silman I, Sussman JL. *Bioinformatics* 2005;21:3435–8. [PubMed: 15955783]
- [28]. Vucetic S, Obradovic Z, Vacic V, Radivojac P, Peng K, Iakoucheva LM, Cortese MS, Lawson JD, Brown CJ, Sikes JG, Newton CD, Dunker AK. *Bioinformatics* 2005;21:137–140. [PubMed: 15310560]
- [29]. Sickmeier M, Hamilton JA, LeGall T, Vacic V, Cortese MS, Tantos A, Szabo B, Tompa P, Chen J, Uversky VN, Obradovic Z, Dunker AK. *Nucleic Acids Res* 2007;35:D786–93. [PubMed: 17145717]
- [30]. Xie Q, Arnold GE, Romero P, Obradovic Z, Garner E, Dunker AK. *Genome Inform Ser Workshop Genome Inform* 1998;9:193–200.
- [31]. Whitley, D. Technical Report CS-93 103 Colorado State University. Fort Collins; CO, USA: 1993. p. 1-39.
- [32]. Michalewicz, Z. *Genetic Algorithms + Data Structures = Evolution Programs*. Springer; Berlin, Germany: 1999. Vol. ' p
- [33]. Vihinen M, Torkkila E, Riikonen P. *Proteins* 1994;19:141–149. [PubMed: 8090708]
- [34]. Cedano J, Aloy P, Perez-Pons JA, Querol E. *J Mol Biol* 1997;266:594–600. [PubMed: 9067612]
- [35]. Fauchere JL, Charton M, Kier LB, Verloop A, Pliska V. *Int J Pept Protein Res* 1988;32:269–78. [PubMed: 3209351]
- [36]. Garbuzynskiy SO, Lobanov MY, Galzitskaya OV. *Protein Sci* 2004;13:2871–7. [PubMed: 15498936]
- [37]. Vacic V, Uversky VN, Dunker AK, Lonardi S. *BMC Bioinformatics* 2007;8:211. [PubMed: 17578581]
- [38]. Anderson, CW.; Appella, E. *Handbook of Cell Signaling*. Bradshaw, RA.; Dennis, EA., editors. Academic Press; New York: 2003. p. 237-247.
- [39]. Palau J, Argos P, Puigdomenech P. *Int J Peptide Protein Res* 1981;19:394–401. [PubMed: 7118409]
- [40]. Nishikawa K, Ooi T. *J Biochem (Tokyo)* 1986;100:1043–7. [PubMed: 3818558]
- [41]. Qian N, Sejnowski TJ. *J Mol Biol* 1988;202:865–84. [PubMed: 3172241]
- [42]. Wimley WC, White SH. *Nat Struct Biol* 1996;3:842–8. [PubMed: 8836100]
- [43]. Nagano K. *J Mol Biol* 1973;75:401–20. [PubMed: 4728695]
- [44]. Krigbaum WR, Komoriya A. *Biochim Biophys Acta* 1979;576:204–48. [PubMed: 760806]
- [45]. Chou PY, Fasman GD. *Adv Enzymol Relat Areas Mol Biol* 1978;47:45–148. [PubMed: 364941]
- [46]. Cheng Y, Oldfield CJ, Meng J, Romero P, Uversky VN, Dunker AK. *Biochemistry* 2007;46:13468–77. [PubMed: 17973494]

Figure 1.

Conditional probability plots calculated for the ordered (black circles) and disordered datasets (open circles) using the following scales: **A.** Normalized flexibility parameters (B-values) for each residue surrounded by two rigid neighbors [33]; **B.** Positive charge [35]; **C.** Composition of amino acids in extracellular proteins [34]; **D.** TOP-IDP.

Figure 2.

Graphical comparison of four scales. **A**. Normalized unfoldability of each residue as assessed by Top-IDP, B-value [33], DisProt and FoldUnfold [36]. Pair-wise comparison of four scales: **B**. TOP-IDP vs. B-value ($r^2 = 0.657$); **C**. TOP-IDP vs. DisProt ($r^2 = 0.609$); **D**. TOP-IDP vs. FoldUnfold ($r^2 = 0.785$); **E**. FoldUnfold vs. DisProt ($r^2 = 0.615$); **F**. FoldUnfold vs. B-value ($r^2 = 0.752$); **G**. DisProt vs. B-value ($r^2 = 0.725$).

Figure 3.
Analysis of the disorder propensity in p53 by Top-IDP (A), PONDRL® VLXT (B) and PONDRL® VSL1 (C).

Table 1

Top 10 scales selected from 517 previously published attributes based on their ARVs

Rank	Description	ARV
1	Normalized flexibility parameters (B-values) for each residue surrounded by one rigid neighbors [33]	0.686942
2	Normalized flexibility parameters (B-values), average [33]	0.675696
3	Normalized frequency of β -sheet from CF [39]	0.672734
4	14 Å contact number [40]	0.672413
5	Weights for β -sheet at the window position of 1 [41]	0.671594
6	Free energies of transfer of AcW1-X-LL peptides from bilayer interface to water [42]	0.670847
7	Normalized frequency of β -structure [43]	0.669485
8	Side chain interaction parameter [44]	0.668985
9	Fraction of site occupied by water [44]	0.667674
10	Normalized frequency of β -sheet [45]	0.666594

Table 2
Comparison of order promoting and disorder promoting amino acids and their associated weights for four scales.

Scale	Order-promoting amino acid residues										
	W	F	Y	I	M	L	V	N	C	T	
Top-IDP	-0.884	-0.697	-0.510	-0.486	-0.397	-0.326	-0.121	0.007	0.02	0.059	
B-value	0.938	0.934	0.981	0.977	0.963	0.982	0.968	1.022	0.939	0.998	
FoldUnfold	28.48	27.18	25.93	25.71	24.82	25.36	23.93	18.49	23.52	19.81	
DisProt	-0.465	-0.381	-0.427	-0.393	0.197	-0.260	-0.302	-0.106	-0.546	-0.116	

Scale	Disorder-promoting amino acid residues										
	A	G	R	D	H	Q	K	S	E	P	
Top-IDP	0.06	0.166	0.180	0.192	0.303	0.318	0.586	0.341	0.736	0.987	
B-value	0.994	1.018	1.026	1.022	0.967	1.041	1.029	1.025	1.052	1.050	
FoldUnfold	19.89	17.11	21.03	17.41	21.72	19.23	18.19	17.67	17.46	17.43	
DisProt	0.042	0.095	0.211	0.127	-0.127	0.381	0.370	0.201	0.469	0.419	