Article

Identifying Similar Patterns of Structural Flexibility in Proteins by Disorder Prediction and Dynamic Programming

Aidan Petrovich [1], Adam Borne [2], Vladimir N. Uversky [3,4,5,6] and Bin Xue [2,*]

[1] Department of Physics, University of South Florida, Tampa, FL 33620, USA;
E-Mail: aidan1@mail.usf.edu

[2] Department of Cell Biology, Microbiology, and Molecular Biology,
School of Natural Sciences and Mathematics, College of Arts and Sciences,
University of South Florida, Tampa, FL 33620, USA; E-Mail: bornea@mail.usf.edu

[3] Department of Molecular Medicine and USF Health Byrd Alzheimer's Research Institute,
Morsani College of Medicine, University of South Florida, Tampa, FL 33620, USA;
E-Mail: vuversky@health.usf.edu

[4] Institute for Biological Instrumentation, Russian Academy of Sciences, Pushchino,
Moscow Region 142290, Russian

[5] Department of Biology, Faculty of Science, King Abdulaziz University, P.O. Box 80203,
Jeddah 21589, Saudi Arabia

[6] Laboratory of Structural Dynamics, Stability and Folding of Proteins, Institute of Cytology,
Russian Academy of Sciences, St. Petersburg 194064, Russian

* Author to whom correspondence should be addressed; E-Mail: binxue@usf.edu;
Tel.: +1-813-974-6007.

Abstract: Computational methods are prevailing in identifying protein intrinsic disorder. The results from predictors are often given as per-residue disorder scores. The scores describe the disorder propensity of amino acids of a protein and can be further represented as a disorder curve. Many proteins share similar patterns in their disorder curves. The similar patterns are often associated with similar functions and evolutionary origins. Therefore, finding and characterizing specific patterns of disorder curves provides a unique and attractive perspective of studying the function of intrinsically disordered proteins. In this study, we developed a new computational tool named IDalign using dynamic programming. This tool is able to identify similar patterns among disorder curves, as well as to present the distribution of

intrinsic disorder in query proteins. The disorder-based information generated by IDalign is significantly different from the information retrieved from classical sequence alignments. This tool can also be used to infer functions of disordered regions and disordered proteins. The web server of IDalign is available at (http://labs.cas.usf.edu/bioinfo/service.html).

Keywords: intrinsic disorder; structural flexibility; disorder pattern; dynamic programming; dynamic time warping

1. Introduction

Computational prediction is now a prevailing strategy in identifying intrinsically disordered proteins (IDPs) and intrinsically disordered regions (IDRs) for both individual proteins and entire proteomes [1–3]. By definition, IDPs/IDRs do not have rigid three-dimensional structures partially due to diminished hydrophobic interactions [4–7] determined by the specific amino acid compositions of IDPs and IDRs, which are typically depleted in hydrophobic, order-promoting residues, but are enriched in polar and charged disorder-promoting residues [8–11]. For this reason, information on the amino acid sequences and compositions of protein chains has been successfully used to predict if a given amino acid or a specific amino acid segment in a query protein is intrinsically disordered or not. Currently, the prediction accuracy of protein intrinsic disorder is reaching ~80% [1], which is becoming comparable with the accuracy of many low-resolution experimental techniques.

Although without rigid structures, IDPs/IDRs are highly abundant in nature and have critical biological functions. It was estimated that the fraction of IDPs/IDRs increases from ~20% in prokaryotic proteome to ~60% in eukaryotic proteomes [12–14]. Some prokaryotic proteomes also have higher fraction of IDPs/IDRs and were found to be associated with the extreme environmental conditions, at which the organisms thrive well [15,16]. The abundance of IDPs/IDRs in various proteomes is a strong indication that these proteins or regions perform important biological functions. Generally speaking, IDPs/IDRs play crucial roles in the processes of cell signaling and regulation [4,17,18]. In addition, IDPs/IDRs are also involved in many other functions. Post-translational modification sites are frequently within or near the IDRs [14,19–28]. Alternative splicing sites are also associated with IDRs [29–32]. Many long IDRs contain short hydrophobic-prone segments. These segments may undergo a disorder-to-order transition when forming complexes with specific binding partners [33]. The binding affinity with partners can also be tuned by IDRs [34,35]. IDPs/IDRs perform their functions through multiple mechanisms. In many cases, functionality is determined by specific amino acid sequences and compositions. In these cases, these functional regions can be recognized by sequence alignments. In other cases, the dynamics or the increased flexibility of disordered residues or regions is also critical. For example, disorder and related high structural flexibility facilitate the process of molecular recognition [36,37]. The dynamics of IDPs/IDRs is also a major contributor to the fuzziness of protein complexes [38].

Whether an amino acid residue or a protein segment is disordered or not can be evaluated by disorder scores generated by specific predictors. The predicted disorder scores of all residues in a query protein can be presented as a curve using the sequential index of amino acid as *x*-axis and disorder

score as *y*-axis. The resulting curve is normally called disorder profile, per-residue disorder plot, or disorder curve. Disorder curve can be used to not only grade the structural flexibility of protein or its segments, but also infer functional roles of IDRs. For example, a "dip" within a segment that have high disorder scores indicates a short structure-prone motif in the middle of long disordered region, and this structure-prone motif may often act as a binding motif, such as molecular recognition feature (MoRF) [39,40] and ANCHOR-identified binding site (AIBS) [41]. Based on these observations, several computational tools have been developed to predict these disorder-based binding motifs [40–44].

The disorder curves of many proteins are similar to each other [45–48]. In addition, as evidenced by some of these studies, the similarity among disorder curves of query proteins often reflects common mechanisms of their function and evolution. The patterns of disordered curves are different from sequential patterns generated by traditional sequence alignment algorithms [49]. The disorder curves provide information on the structural flexibility, which may hardly be inferred from sequences directly. In a recent study, the peculiarities of disorder pattern were found to be critical for ion binding and the functions of heparinase II in *Pedobacter heparinus* [50]. In another recent study, comparison of disorder profiles was used to analyze regions that are involved in isoform-specific binding of tropomodulin (Tmod) to tropomyosin (TM), and to predict the residues that characterize isoform differences in binding [34]. In another related study, comparative analysis of disorder profiles of the wild type and the mutant forms of Tmod-1 and the wild type of Tmod-4 was used to define mutations that would affect the affinity of Tmod-1 to skeletal striated TM and make it similar to that of Tmod-4 [51]. A recent analysis on a dormancy-associated plant gene family DORMANCY 1/AUXIN Represses Protein (DRM1/ARP) revealed that these plant proteins can be grouped into six distinct classes based upon the similarity of their disorder profiles [52]. Similar analysis of another group of plant proteins, RPM1-interacting proteins 4 (RIN4) that belongs to the family of proteins containing nitrate-induced (NOI) domains and playing important roles in the plant immune responses to various pathogens, provided another proof that comparison of disorder curves facilitates functional annotation of proteins [53]. Additionally, disorder-based sequence alignments were used to show similarity of disorder distribution in several milk proteins, such as different casein classes [54], lactoperoxidases [55], and C- and N-lobes of lactoferrin [56]. Additionally, a new concept of *de novo* design of artificial IDPs was also recently brought into the scientific community [57]. Apparently, characterizing disorder patterns is a prerequisite for these new advancements in the IDP field.

However, to the best of our knowledge, none of the current computational tools in this field is specifically designed for comparative analysis of disorder patterns. Therefore, to fill this gap, we developed a novel computational tool to measure the similarity of different disorder curves by using dynamic programming. Dynamic programming has been broadly used in time series data analysis [58], sequence alignment [59], and string match [60]. In our study, dynamic programming was for the first time applied to compare disorder curves. It is expected that the results from this study provide new ideas to characterize patterns of intrinsic disorder and to infer functions associated with structural flexibility.

2. Results

2.1. Building up the Dataset

Figure 1 shows the distribution of proteins on a two-dimensional space formed by the length of protein sequences and the fraction of disordered residues in the sequence, for yeast proteome and for all disordered proteins in DisProt. Although the actual number of proteins is very different (6660 in yeast and 694 in DisProt), the overall distributions are similar. Most proteins have less than 500 residues, and less than 30% of disordered residues. It is also obvious that a small group of proteins in DisProt, which have less than 200 residues, is characterized by a high fraction of disordered residues (~100%). For the computational efficiency, 100 sequences shorter than 400 residues were randomly selected from each of these two sets, yeast proteome and DisProt proteins. These two dataset are referred to as Y100 and D100 dataset, respectively.
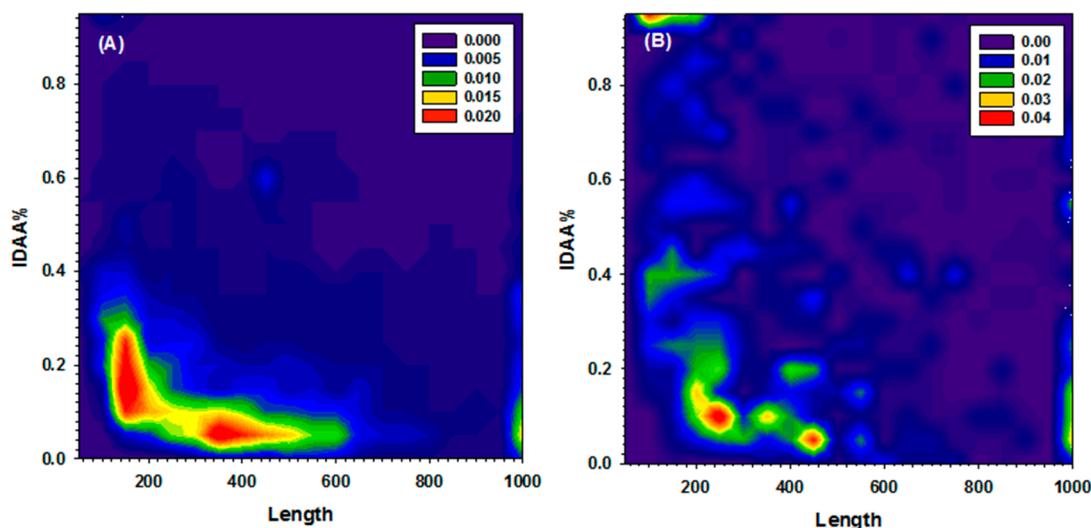

Figure 1. Abundance of proteins as a function of length and fraction of disordered residues (IDAA%) in both (**A**) Yeast and (**B**) DisProt datasets. All protein sequences longer than 1000 residues were merged with into the group of proteins of 1000 residues. The per-residue disorder score was calculated from PONDR-FIT. All residues of which the disorder score is higher than 0.5 were counted as disordered residues. The fraction of disordered residues is the ratio over the length of corresponding protein. Colors from purple, to blue, green, yellow, and red represents the increased abundance.

2.2. Gap Penalty

For all sequence pairs in each of the datasets, their alignment scores increased when the gap penalty increased as shown in Figure 2. Nonetheless, the increment of alignment scores saturated after reaching the threshold values, indicating that the alignment is becoming stable. Therefore, the mean value of the normalized alignment scores over all sequence pairs in a dataset was used in Figure 2 to find the optimal threshold value of gap penalty for that dataset. The optimized gap penalty values for Y100 dataset and D100 dataset are slightly different, with the former being 0.3 and the latter being 0.5. In addition, mean value of normalized fraction of matches was also calculated to compare with the

alignment score. It should be noted that all sequence pairs in each dataset can be divided into two sub-groups: (1) sequence pairs for which the fraction of matches decreases with the raising penalty score (typically, these are pairs with very low sequence similarities); and (2) sequence pairs for which the fraction of matches increases with the penalty score. Figure 2 demonstrates the presence of these two types of sequence pairs by showing two opposite trends in the correlation between fraction of matches and gap penalty scores. The lower standard errors within each group further validated the consistence of sequences within each group. This evidence added a second requirement on the selection of threshold value: The threshold value of gap penalty should have an intermediate value to balance the opposite influence on two sub-groups of sequence pairs. Therefore, after taking into consideration of all the factors, we chose 0.4 as the optimized gap penalty in all further studies.
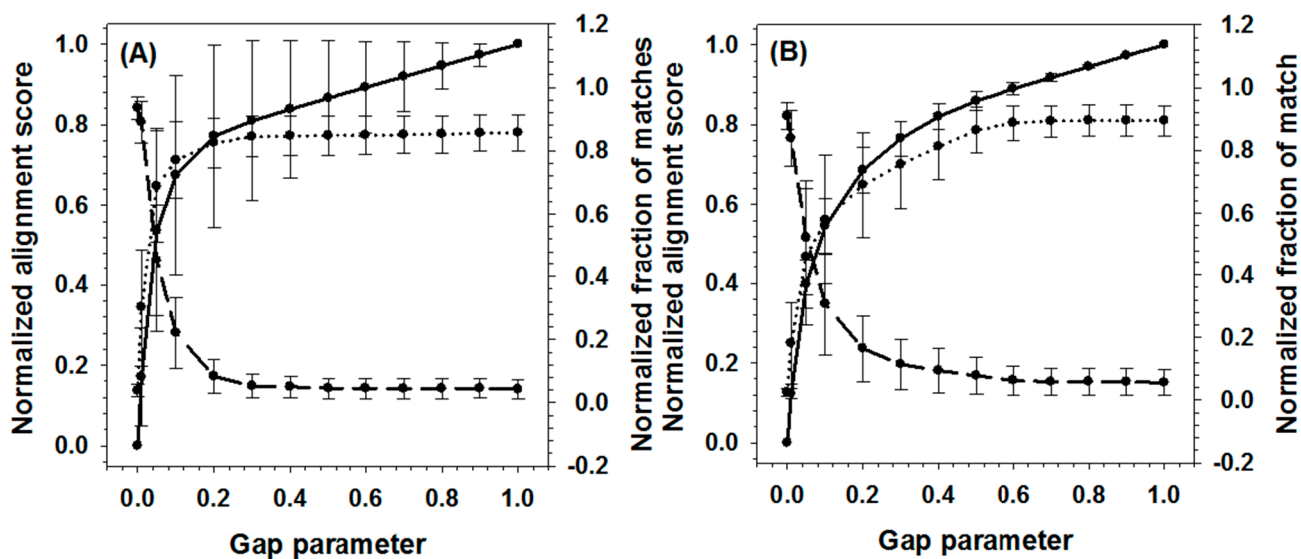

Figure 2. Influence of gap parameter on global matches for (**A**) Yeast dataset and (**B**) DisProt dataset. The solid line shows for normalized alignment score, while dash line and dotted lines represent normalized fraction of matches for two types of sequence pairs. Error bars present standard error. The first type of sequence pairs has lower similarity on their disorder curve and the calculated fraction of matches increases with gap penalty. The second type of sequence pairs is on the contrary. They have higher similarity on their disorder curves and their calculated fraction of matches decreases with the gap penalty.

2.3. Match Threshold

After the alignment path being identified, the matches between data points were determined by comparing their pair-wise distance with the threshold of match score V_{match}. The value of V_{match} will not affect the final alignment score, but only the fraction of identified matches. Conceivably, the larger the V_{match}, the higher the fraction of matches. Figure 3 presented the analysis on the correlation between fraction of matches and V_{match} for both Y100 and D100 datasets. At the low end of each match threshold, the fractions of matches increased rapidly with the threshold value in both datasets. After reaching 0.05, the fractions became stabilized. Therefore, with the purpose of limiting the number of matched segments identified in the alignment, 0.05 was used as the threshold value for the matched data points in the application.
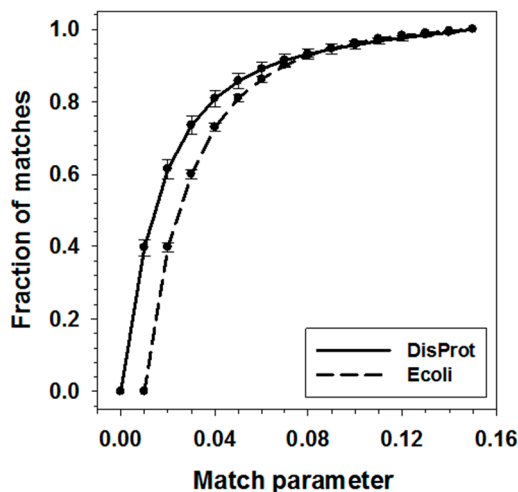
Figure 3. Match parameter influences fraction of identified matches. The averaged fraction of matches on one pair of sequences was calculated and then normalized in the datasets. Solid line and dash line are the correlation between match parameter and fraction of matches for DisProt and Yeast datasets, respectively.

2.4. Examples/Applications

To examine the powerfulness and usefulness of this newly developed computational strategy, we performed several case studies. Figure 4 presents the similarity of disorder curves between two uncharacterized proteins from yeast (A0A023PXP4 and A0A023PZE6). These two proteins have different lengths, rather different sequences (see Figure 4A), and different fractions of disordered residues. However, they do have local similarity in their disorder profiles, such as the presence of the double-peak segments in the middle of their sequences, and the specific "dips" within their C-terminal tails (Figure 4B). The contour map in Figure 4C describes the similarity in more details. In the contour map, darker colors represent lower distance scores and therefore more similarity. Therefore, the region filled by darker colors connecting the first residue to the last residue tracks down a warping path and therefore represents the alignment path. It is clear that at the *N*-terminal ends, the alignment path is off-diagonal. In the region from 20 to 60 on *x*-axis and from 40 to 80 on *y*-axis, the alignment path becomes diagonal, representing matched curves. This region corresponds to the double-peak segments on both disorder curves. Afterwards, following another off-diagonal segment, the alignment path becomes narrow and diagonal in the range from ~110 to ~130 on *x*-axis and ~90 to ~100 on *y*-axis, indicating a highly matched curves at the C-terminal tails. By using the outputs from the identified alignment path, the original disorder curves were stretched and aligned in Figure 4D. The highlighted four short regions were characterized by similar patterns of the disorder distribution. By comparing the results of sequence alignment in Figure 4A, it is clear that these regions with matching disorder profiles have very limited sequence similarity.

(A)

```
A0A023PXP4    MLIGAPSNMRLRGALELLWRRLLHGLMQLRLVLKMHICSQLNHAIKRRAFLKTKGGKNAL
A0A023PZE6    -------MSSLLQLLAVWS-------QSSSISMNSEVAQTNQKYKDHSFLVSQSFYSPF
                 *        *  :*            *    :        :* *: * ::** ::. ..:

A0A023PXP4    HG-CLDTVINLQESILAGIRLVPVLPILLDYVLYNISLGGMTFADFLEVLLHFSALEGLC
A0A023PZE6    VNSCYWKLVNRRADLLKILFLRRKGDMHSHLFSYTFIFY---FLLFKILSLHPFREPAKI
                . *  .::* :  .:*   :  *        :  . . *.: :      *  *  : **        .

A0A023PXP4    KGVFEADGLEAVH
A0A023PZE6    NI-----------
                :
```


Figure 4. Identified alignment path for a sequence pair (Uniprot IDs: A0A023PXP4 and A0A023PZE6) from the Yeast dataset. (**A**) Traditional pair-wise sequence alignment. "*****", "**:**", "**.**", and "**-**" stand for identical amino acids, highly similar amino acids, similar amino acids, and gaps, respectively; (**B**) Original disorder predictions for A0A023PXP4 (**upper** panel) and A0A023PZE6 (**lower** panel). The gray shadow behind the disorder curves is estimated prediction error from PONDR-FIT; (**C**) Alignment path between two sequences identified by our newly developed package; (**D**) Alignment of disorder curves between A0A023PXP4 (pink) and A0A023PZE6 (black) along the alignment path. Many pairs of segments between pink and black curves overlap with each other. Only the pairs, of which the distance between two segments less than 0.05, were highlighted by cyan.

Another example for the alignment of disorder curves between DP00270 (anti-ssDNA Fab DNA-1) and DP00710 (Fab fragment of immunoglobulin G1 MAK33 heavy chain) from the DisProt database is presented in Figure 5. These two proteins have very similar disorder curves in their C-terminal tails (Figure 5B). Actually, as shown in Figure 5A, the C-terminal tails of these two proteins have almost identical sequences. The alignment path in Figure 5C is almost completely diagonal but the color is lighter than that in Figure 4C. Multiple matched regions were identified in Figure 5D. Although the sequences become nearly identical starting at around residue 100, the disordered curves have less matches. The discrepancy comes from the predictive results of IUPred and will be explored further in the discussion section. An important conclusion from examples shown in both Figures 4 and 5 is that the alignment of disorder curves provides additional information that may not be revealed by traditional sequence alignment.

At the next stage, we analyzed the usefulness of IDalign for the analysis of an important protein p53 and its homologues. Although this tumor suppressor is a well-known protein that does not require long introduction, some important information is provided below. The activity of this crucial transcription factor is modulated by various stress signals affecting genome integrity and cell proliferation.

Activation of p53 triggers a complex cellular response regulating expression of genes involved in various biological processes, such as DNA repair, cell cycle progression, induction of apoptosis, response to cellular stress, senescence, *etc.* [61–63]. Some developmental abnormalities in animals are associated with the p53 deficiency [64]. Furthermore, the loss of p53 function is often related to the cancerous transformation of the cell [65]. In fact, cancers showing mutations in p53 are found in colon, lung, esophagus, breast, liver, brain, and in hemopoietic and reticuloendothelial tissues [65]. Human p53 is a 393 residue-long protein containing three functional regions, the N-terminal region, the central DNA Binding Domain (DBD), and the C-terminal region [62]. The N-terminal region can be further subdivided into TransActivation Domain 1 (TAD1) (residues 1–40), TAD2 (residues 40–60), and a Proline-Rich region, PR (residues 64–92). The C-terminal region contains a tetramerization or Oligomerization Domain (OD; residues 325–356), and a regulatory C-Terminal Domain (CTD; residues 356–393) [62,66]. Intrinsic disorder is known to be crucial for function of p53 [67–69], where, for example, the intrinsically disordered C-terminal region possesses a unique binding plasticity, being able not only to interact with various binding partners, but also to gain different structures in its bound form [70].
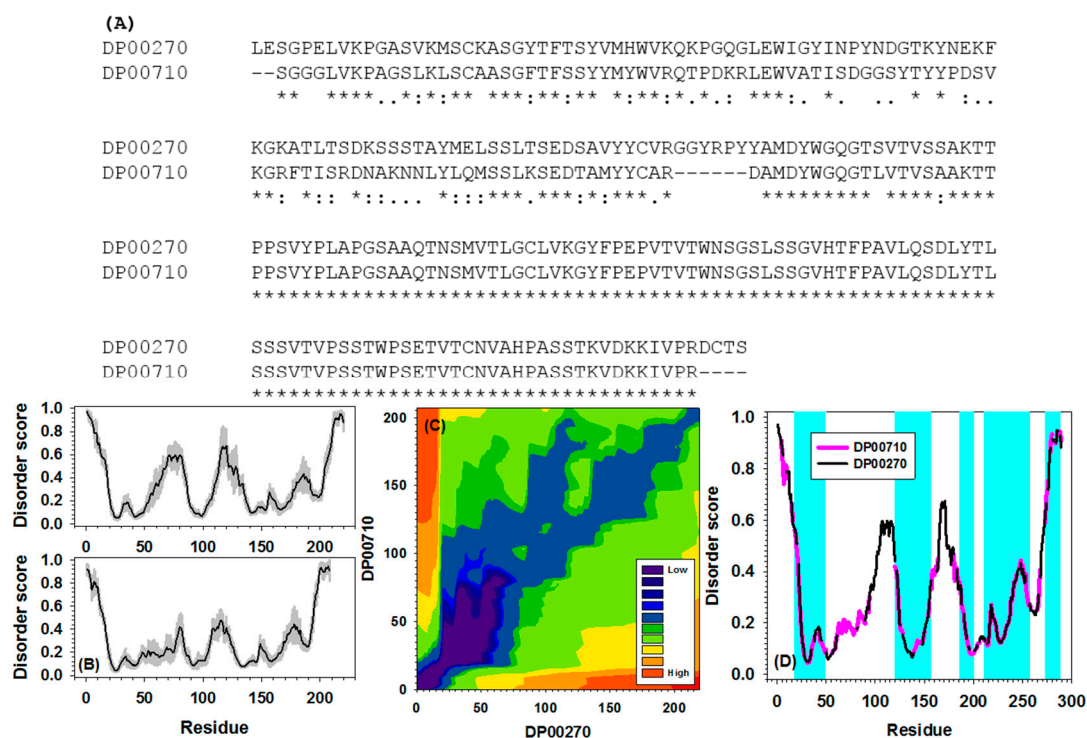
Figure 5. Identified alignment path for a sequence pair (Disprot IDs: DP00270 and DP00710) from the DisProt dataset. (**A**) Traditional pair-wise sequence alignment. "*", ":", ".", and "-" stand for identical amino acids, highly similar amino acids, similar amino acids, and gaps, respectively; (**B**) Original disorder prediction for DP00270 (**upper** panel) and DP00710 (**lower** panel). The gray shadow behind the disorder curves is estimated prediction error; (**C**) Alignment path between two sequences identified by our newly developed package; (**D**) Alignment of disorder curves between DP00710 (pink) and DP00270 (black) along the alignment path. Only overlapped segment pairs of which the distance between two segments lower than 0.05 were highlighted by cyan.

For human p53, disorder evaluations together with important disorder-related functional information were retrieved from D^2P^2 database (http://d2p2.pro/) [71]. D^2P^2 is a database of predicted disorder that represents a community resource for pre-computed disorder predictions on a large library of proteins from completely sequenced genomes [71]. D^2P^2 database uses outputs of PONDR® VLXT [8], IUPred [72], PONDR® VSL2B [73,74], PrDOS [75], ESpritz [76], and PV2 [71]. This database is further enhanced by information on the curated sites of various posttranslational modifications and on the location of predicted disorder-based potential binding sites. Figure 6 represents the results of the application of this tool to human p53 and provides further support for the abundance and functional importance of intrinsic disorder in this protein. In fact, Figure 6 shows that this protein contains long disordered regions, which are enriched in potential disorder-based binding motifs and numerous sites of posttranslational modifications, PTMs. The fact that disordered domains/regions of human p53 are heavily enriched in various PTM sites is in agreement with the well-known notion that phosphorylation [77] and many other enzymatically catalyzed PTMs are preferentially located within the IDPRs [28].
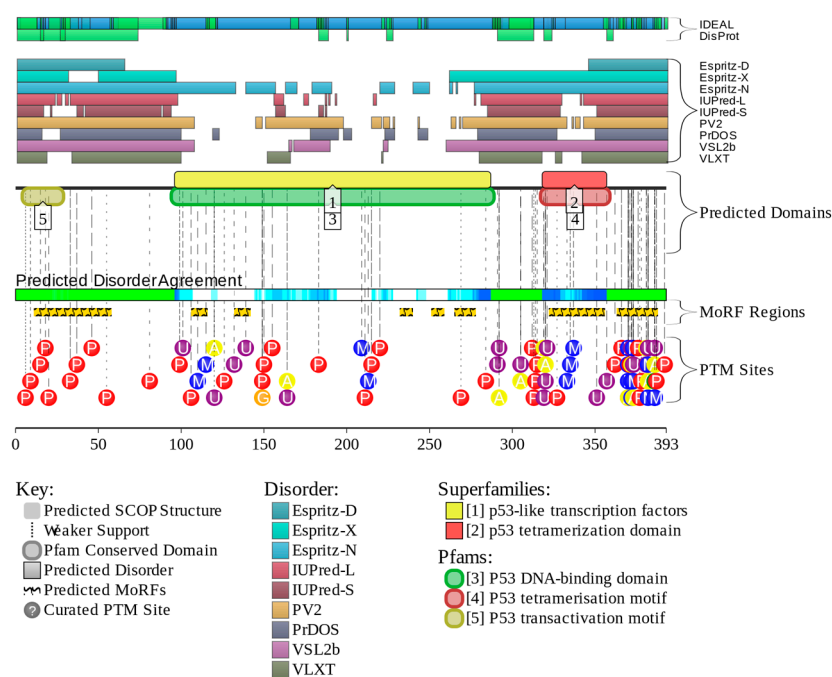
Figure 6. Evaluation of the functional intrinsic disorder propensity of human p53 (UniProt ID: P04637) by D^2P^2 database (http://d2p2.pro/) [71]. In this plot, top two lines represent annotated disordered regions in the DisProt and IDEAL databases. Next nine colored bars represent location of disordered regions predicted by different disorder predictors (Espritz-D, Espritz-N, Espritz-X, IUPred-L, IUPred-S, PV2, PrDOS, PONDR® VSL2b, and PONDR® VLXT, see keys for the corresponding color codes). Green-and-white bar in the middle of the plot shows the predicted disorder agreement between these nine predictors, with green parts corresponding to disordered regions by consensus. Yellow bar shows the location of the predicted disorder-based binding site (MoRF region), whereas colored circles at the bottom of the plot show location of sites of various posttranslational modifications (red—phosphorylation, blue—methylation, yellow—acetylation; orange—glycosylation; and violet—ubiquitylation).

Figure 7 represents the results of the IDalign-based alignments of the disorder profiles of human p53 with its evolutionary distant homologues, p53 proteins from fish (UniProt ID: P79820) and fly (UniProt ID: Q9N6D8). In both cases, the resulting contour maps (Figure 7A,C), especially within the N-terminal regions of corresponding pairs, are asymmetric, with regions with darker colors that correspond to more similar sequence segments being located off-diagonal. Figure 7A,C also shows that there is another level of asymmetry, since the disorder-based similarity worsens while moving to the N- to the C-terminus, giving rise to the dark, off-diagonal, N-terminal regions and noticeably lighter, mostly on-diagonal, C-terminal regions. By using the outputs from the identified alignment paths, the original disorder curves were stretched and aligned (see Figure 7B,D). The highlighted short regions in Figure 7B,D correspond to sequence segments characterized by similar patterns of the disorder distribution. As expected, the number of these similar patterns is lower in more distant human-fly pair. The corresponding traditional sequence alignments are shown in Figure S1.
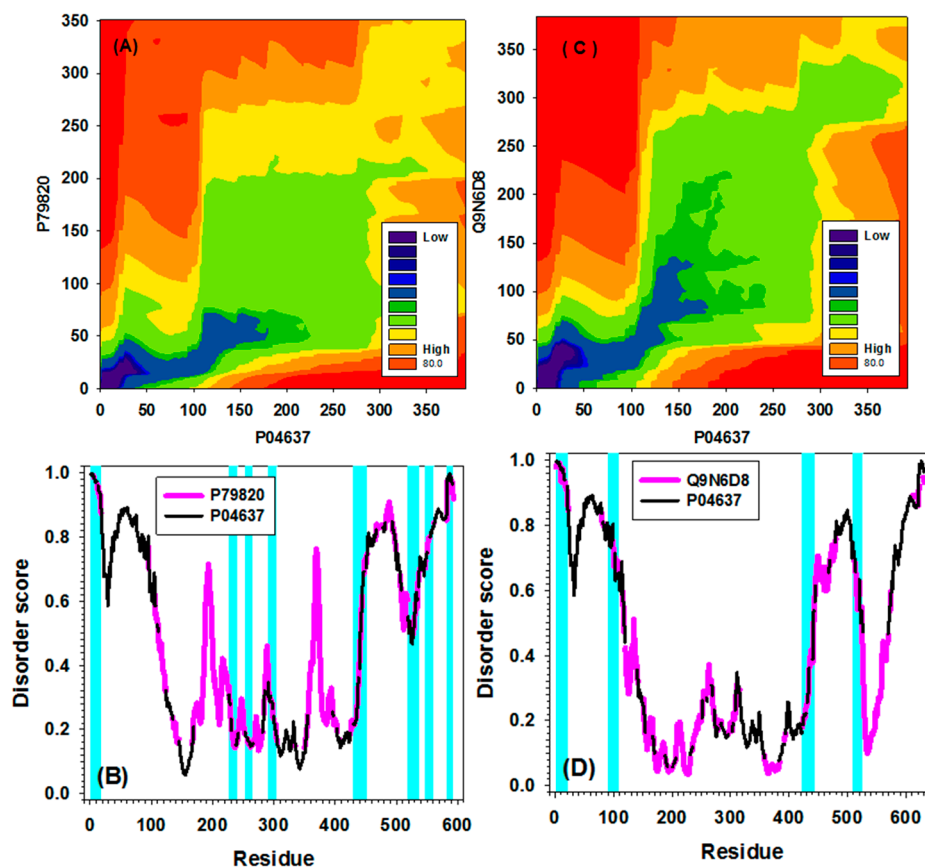
Figure 7. Identified alignment paths and alignments for sequence pairs between human p53 (Uniprot ID: P04637) and fish p53 (Uniprot ID: P79820) in (**A**,**B**), and between human p53 and fly p53 (Uniprot ID: Q9N6D8) in (**C**,**D**), respectively. (**A**,**C**) Alignment paths (contour maps) between two sequences in each of the sequence pairs were identified using our newly developed package; (**B**,**D**) Alignment of disorder curves along the alignment paths for two sequence pairs: P79820 (pink) and P04637 (black) in (**B**); Q9N6D8 (pink) and P04637 (black) in (**D**). Only overlapped segments of which the distance less than 0.05 were highlighted by cyan.

The p53 protein is a member of an important protein family that includes p53, p63 (see [78,79]) and p73 [80]. Both p63 and p73 are structurally similar and functionally related to p53 [67]. The members of the p53 family are interlinked in a unique family-based signaling network that controls various aspects of cell life such as proliferation, differentiation, and death [63]. Both human p63 and p73 are almost two-fold longer than p53 and have 680 and 636 residues respectively. The domain organization of the members of p53-family is rather similar, and all three proteins have identifiable TAD, DBD and OD. In p63 and p73, there is an additional C-terminal sterile-α motif (SAM), which is required for p63 and p73 transcriptional activity, but they seem to lack the CTD found in p53 [63]. The various p53 family members have limited overall homology, but strong similarity in the DBD (approximately 60% between p53 and p63/p73 and approximately 85% between p63 and p73) [81,82]. Figure 8A,C represents the contour plots and the aligned profiles of human p53-p63 and p63-p73 pairs. Here, likely due to the large difference in the sequence lengths, the p53-p63 alignment is highly asymmetric, with the vast majority of darker regions being located off-diagonal (see Figure 8A). Although the contour plot representing the p63-p73 alignment is more symmetric (Figure 8C), this plot is characterized by noticeably lighter colors, which is expected due to the limited overall sequence homology of these two proteins. Figure 8B,D represents stretched and aligned disorder profiles of the human p53-p63 and p63-p73 pairs respectively. The highlighted short regions in Figure 8B,D correspond to sequence segments characterized by similar patterns of the disorder distribution. Note that these highlighted regions are concentrated mostly around the DBDs, reflecting higher levels of sequence/disorder pattern similarity in these domains in comparison with other regions. Again, the corresponding traditional pair-wise sequence alignments for human p53-p63 and p63-p73 pairs are shown in Figure S1.

2.5. Web Server

In order to facilitate the large-scale proteomic analysis of the common patterns of disorder curves, as well as the accurate positioning of the matched segments, we further developed a webserver, which can be accessed through the following link (http://labs.cas.usf.edu/bioinfo/service.html). The layout of this web server is shown in Figure 9. Users may choose one of the following two methods to input data. In the first method, the users may input comma-delimited disorder scores for two proteins. In the other method, the users may upload two data files that contain disorder scores in a single-column format. The output of the web server has three columns. The first column is the sequential index after alignment. The second and the third columns are disorder scores after alignment for the 1st and 2nd sets of input data, respectively. It should be noted that if two curves do not match to each other at a specific position, the score will be assigned as "−1".
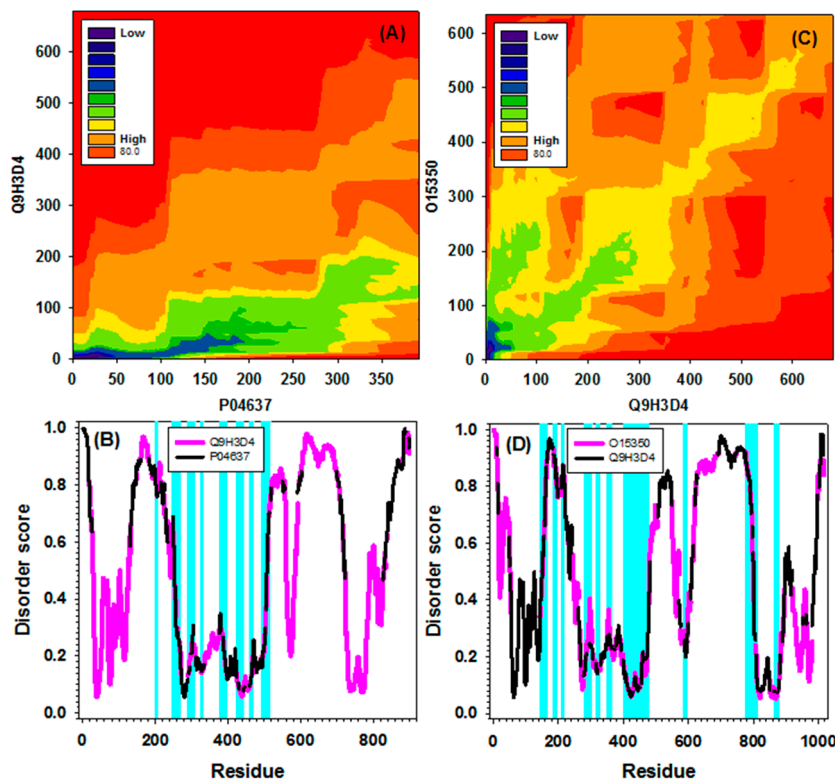
Figure 8. Identified alignment paths and alignments for sequence pairs between human p53 (Uniprot ID: P04637) and human p63 (Uniprot ID: Q9H3D4) in (**A**,**B**), and between human p63 and human p73 (Uniprot ID: O15350) in (**C**,**D**), respectively. (**A**,**C**) Alignment paths (contour maps) between two sequences in each of the sequence pairs were identified by our newly developed package; (**B**,**D**) Alignment of disorder curves along the alignment paths for two sequence pairs: Q9H3D4 (pink) and P04637 (black) in (**B**); O15350 (pink) and Q9H3D4 (black) in (**D**). Only overlapped segment pairs of which the distance less than 0.05 were highlighted by cyan.
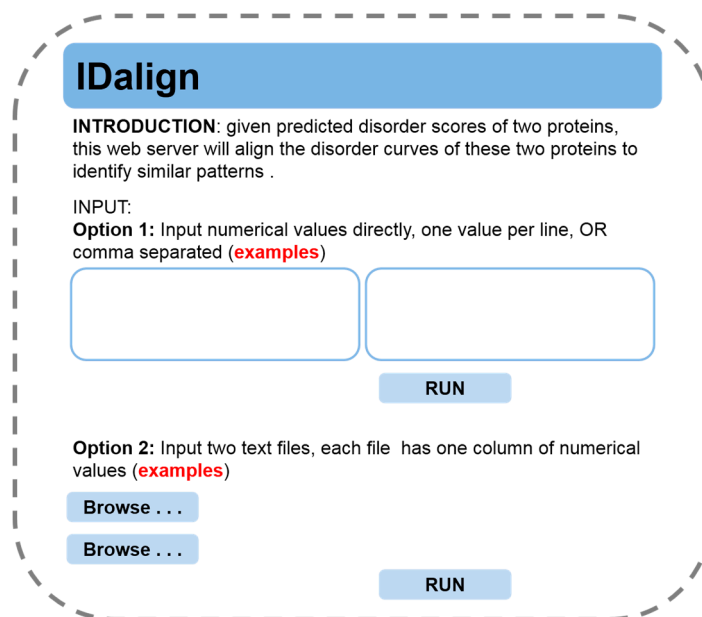
Figure 9. Layout the IDalign web server.

3. Discussions

It has been well recognized in the field of dynamic programming that both the gap penalty and the calculation of cost matrix may significantly shift the identified global matches. In various applications of dynamic time warping algorithms, the calculation of distance and cost matrix is extremely critical for the applicability of the algorithm. In this study, the calculation of distance is rather straight forward due to the nature of the problem. However, we did observe the significant influence of the gap penalty on the final output. When the gap penalty is low, a small increment may change the alignment score remarkably, as well as change the fraction of matches.

In addition, we observed the presence of two types of sequence pairs in both yeast and DisProt datasets. One type of the sequence pairs had low similarity of the corresponding disorder curves, and their overall fraction of matches increased with the gap penalty. The other type of sequence pairs had higher similarity of their disorder curves, and their fraction of matches decreased when the value of gap penalty was raised.

The aforementioned observations imposed rather strict criteria for the selection of gap penalties. Alignment score became saturated at higher values of gap penalty, indicating a stable alignment. In this meaning, the larger the gap penalty, the better the alignment. However, due to the presence of two types of sequence pairs, of which one type has decreased fraction of matches at higher penalty and the other has increased fraction of matches, a gap penalty of intermediate values may suit better to a general purpose alignment.

The disorder curves used in this study were generated by the PONDR-FIT, which is a meta-predictor built on six component predictors: PONDR-VLXT, PONDR-VSL2, PONDR-VL3, IUPred, FoldIndex, and TopIDP [83]. Most of the component predictors applied the sliding-window technique to take into consideration the influence of neighboring residues on the disorder score of the query residue, which is in the central of the sliding-window. Therefore, the disorder score of one residue is determined by all other residues inside the sliding-window that is normally 20 to 30 residues. In this respect, the IUPred [72] is very different from other component predictors. IUPred applies the pairwise amino acid interactions from next 100 amino acids along the sequence. Therefore, the calculated per-residue score from IUPred may be affected by the sequentially more distant residues. Consequently, the PONDR-FIT score may also be affected by the residues far away from the query residue. That is the reason why highly identical sequences in Figure 5D still have obviously different patterns of their disorder curves. To further validate our argument for the influence of IUpred on the alignment path, we tested the alignment between DP00710 and DP00270 using PONDR@VLXT scores [8] (Figure S2), which does not include the influence of long-range amino acids. It is clear that in the alignment of PONDR@VLXT scores, the influence of long-range amino acids is no longer present. Nonetheless, this difference shown in Figure 5D is advantageous since the result from IUPred takes into consideration the presence of long-range interactions, which are missed in other sliding-window methods. Therefore, the disorder curve from PONDR-FIT, which incorporates the results of IUPred, reflects partially the influence of long-range interactions on structural flexibility. That is also the reason why disorder-curve-based alignment provides additional information that is overlooked by sequence-alignment based methods.

4. Experimental Section

4.1. Datasets and Disorder Prediction

The entire proteome of *Saccharomyces cerevisiae* (strain ATCC 204508/S288c) was downloaded from UniProt (ftp://ftp.uniprot.org/pub/databases/uniprot/). This proteome contains 6740 protein sequences. After removing sequences shorter than 40 residues, the remaining dataset has 6660 proteins. All these protein sequences were predicted using PONDR-FIT [83] to get the per-residue disorder scores. In addition, the fraction of disordered residues was calculated for each sequence using 0.5 as threshold value for disordered residues. Finally, a subset of proteins of which the lengths are between 100 and 400, and the fractions of disordered residues are between 10% and 30% was randomly selected. The reasons for filtering out sequences shorter than 100 and longer than 400 amino acids are: (1) the prediction accuracies of terminal residues are normally lower than the accuracy of internal residues [84]. Since many predictors use sliding windows of 20~30 amino acids, the accuracies on the first and the last 20~30 residues in the sequence will be affected. Therefore, we chose 100 amino acids as the lower limit of the length of sequences in the dataset to ensure the overall acceptable prediction accuracy; (2) proteins are often organized in different domains. The length of a single domain is normally from tens of residues to around 400 amino acids. In other words, sequences longer than 400 amino acids may be composed of multiple domains and contain many linker regions. Therefore, we selected sequences shorter than 400 amino acids to build the dataset. The final dataset includes 100 sequences and is addressed as the Y100 dataset.

The second dataset used in this study is a subset of experimentally validated disordered proteins from DisProt [85]. DisProt has a total of 694 disordered protein sequences. All the sequences were also predicted by PONDR-FIT. Afterwards, 100 sequences with length in the range from 100 to 400 residues were selected to compose the D100 dataset.

4.2. Dynamic Programming

Assume the first disorder curve has N data points (x_1, x_2, . . ., x_N), and the second one has M data points (y_1, y_2, . . ., y_M). The purpose of this study is to calculate the pair-wise distance between the data points on two different curves and finally to evaluate the similarity between two curves. We designed an algorithm similar to dynamic time warping (DTW) [86] to compare the disorder curves. The Euclidian distance between any two data points from two different curves was calculated as $f_{i,j} = |x_i - y_j|$, I = (1, N) and j = (1, M). The cost function in the algorithm is $F_{i,j} = \min (f_{i-1, j-1}, f_{i-1, j}, f_{i, j-1}) + f_{i, j}$. In addition, we also introduced a gap penalty score P when initializing the cost function. The gap penalty serves as a global constraint and does influence the results of identified global matches. The pseudo code for the implementation of the algorithm is shown in Figure 10.
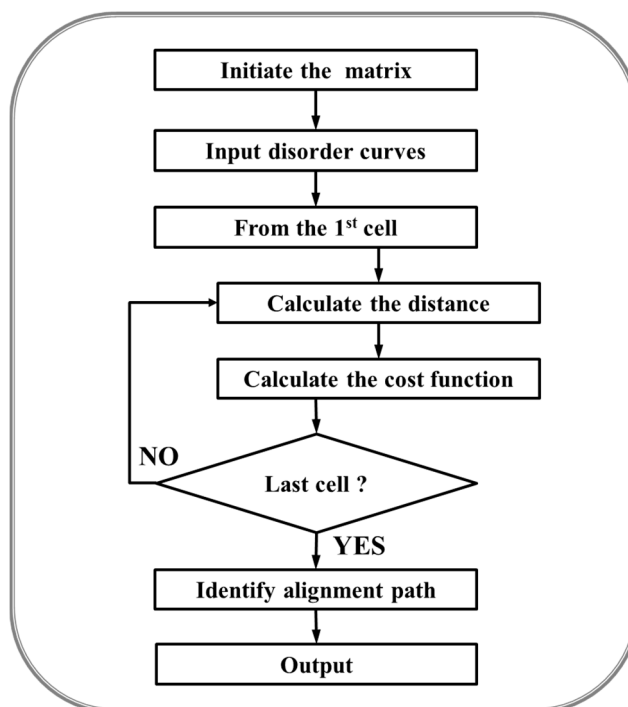
Figure 10. Pseudo code of the algorithm. Dynamic programming was applied to search for the similarity between two disorder curves. When initiating the matrix, the penalty score P was assigned to the first column and first row of the matrix. Then the data points on the disorder curves were uploaded into the 2nd column and 2nd row of the matrix. Next, the distance and cost function were calculated using the formula described in the method section starting from the first vacant cell. After completing the calculation for all cells in the matrix, the alignment path was identified starting from the last cell to the first cell by connecting cells with lower cost function values.

4.3. Fraction of Matches

The data points in this study are per-residue disorder scores, which ranges from 0 to 1. The result from above-mentioned dynamic programming analysis is an alignment path (or warping path as used in DTW). Therefore, the distances between data points from two curves along this path can be calculated. The distances can also be compared to a threshold value V_{match}. If the distance between two data points is less than V_{match}, these two data points are considered to be a pair of matches. Furthermore, the fraction of matched data points for a pair of curves can be calculated as $f_M = 2 \times N_{match}/(N_1 + N_2)$. N_{match} is the number of matched data points. N_1 and N_2 are the lengths of two curves.

Supplementary Materials

Supplementary materials can be found at http://www.mdpi.com/1422-0067/16/06/13829/s1.

Author Contributions

Bin Xue designed the experiment, Bin Xue and Aidan Petrovich developed the algorithm and the computational code, Bin Xue and Aidan Petrovich collected and analyzed the data, Adam Borne and Bin Xue designed the web server, Bin Xue, Aidan Petrovich and Vladimir N. Uversky wrote the manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Monastyrskyy, B.; Kryshtafovych, A.; Moult, J.; Tramontano, A.; Fidelis, K. Assessment of protein disorder region predictions in casp10. *Proteins* **2014**, *82* (Suppl. 2), 127–137.
2. Ali, H.; Urolagin, S.; Gurarslan, O.; Vihinen, M. Performance of protein disorder prediction programs on amino acid substitutions. *Hum. Mutat.* **2014**, *35*, 794–804.
3. Punta, M.; Simon, I.; Dosztanyi, Z. Prediction and analysis of intrinsically disordered proteins. *Methods Mol. Biol.* **2015**, *1261*, 35–59.
4. Wright, P.E.; Dyson, H.J. Intrinsically unstructured proteins: Re-assessing the protein structure-function paradigm. *J. Mol. Biol.* **1999**, *293*, 321–331.
5. Dunker, A.K.; Lawson, J.D.; Brown, C.J.; Williams, R.M.; Romero, P.; Oh, J.S.; Oldfield, C.J.; Campen, A.M.; Ratliff, C.M.; Hipps, K.W.; *et al.* Intrinsically disordered protein. *J. Mol. Graph. Model.* **2001**, *19*, 26–59.
6. Uversky, V.N.; Gillespie, J.R.; Fink, A.L. Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins* **2000**, *41*, 415–427.
7. Tompa, P. Intrinsically unstructured proteins. *Trends Biochem. Sci.* **2002**, *27*, 527–533.
8. Romero, P.; Obradovic, Z.; Li, X.; Garner, E.C.; Brown, C.J.; Dunker, A.K. Sequence complexity of disordered protein. *Proteins* **2001**, *42*, 38–48.
9. Adkins, J.N.; Lumb, K.J. Intrinsic structural disorder and sequence features of the cell cycle inhibitor p57kip2. *Proteins* **2002**, *46*, 1–7.
10. Weathers, E.A.; Paulaitis, M.E.; Woolf, T.B.; Hoh, J.H. Reduced amino acid alphabet is sufficient to accurately recognize intrinsically disordered protein. *FEBS Lett.* **2004**, *576*, 348–352.
11. Hansen, J.C.; Lu, X.; Ross, E.D.; Woody, R.W. Intrinsic protein disorder, amino acid composition, and histone terminal domains. *J. Biol. Chem.* **2006**, *281*, 1853–1856.
12. Peng, Z.; Yan, J.; Fan, X.; Mizianty, M.J.; Xue, B.; Wang, K.; Hu, G.; Uversky, V.N.; Kurgan, L. Exceptionally abundant exceptions: Comprehensive characterization of intrinsic disorder in all domains of life. *Cell. Mol. Life Sci.* **2015**, *72*, 137–151.

13. Peng, Z.; Mizianty, M.J.; Kurgan, L. Genome-scale prediction of proteins with long intrinsically disordered regions. *Proteins* **2014**, *82*, 145–158.

14. Xu, K.; Uversky, V.N.; Xue, B. Local flexibility facilitates oxidization of buried methionine residues. *Protein Pept. Lett.* **2012**, *19*, 688–697.

15. Xue, B.; Williams, R.W.; Oldfield, C.J.; Dunker, A.K.; Uversky, V.N. Archaic chaos: Intrinsically disordered proteins in archaea. *BMC Syst. Biol.* **2010**, *4* (Suppl. 1), S1.

16. Pavlovic-Lazetic, G.M.; Mitic, N.S.; Kovacevic, J.J.; Obradovic, Z.; Malkov, S.N.; Beljanski, M.V. Bioinformatics analysis of disordered proteins in prokaryotes. *BMC Bioinform.* **2011**, *12*, 66.

17. Dunker, A.K.; Brown, C.J.; Obradovic, Z. Identification and functions of usefully disordered proteins. *Adv. Protein Chem.* **2002**, *62*, 25–49.

18. Minezaki, Y.; Homma, K.; Kinjo, A.R.; Nishikawa, K. Human transcription factors contain a high fraction of intrinsically disordered regions essential for transcriptional regulation. *J. Mol. Biol.* **2006**, *359*, 1137–1149.

19. Xie, H.; Vucetic, S.; Iakoucheva, L.M.; Oldfield, C.J.; Dunker, A.K.; Obradovic, Z.; Uversky, V.N. Functional anthology of intrinsic disorder. 3. Ligands, post-translational modifications, and diseases associated with intrinsically disordered proteins. *J. Proteome Res.* **2007**, *6*, 1917–1932.

20. Eisenhaber, B.; Eisenhaber, F. Posttranslational modifications and subcellular localization signals: Indicators of sequence regions without inherent 3d structure? *Curr. Protein Pept. Sci.* **2007**, *8*, 197–203.

21. Radivojac, P.; Vacic, V.; Haynes, C.; Cocklin, R.R.; Mohan, A.; Heyen, J.W.; Goebl, M.G.; Iakoucheva, L.M. Identification, analysis, and prediction of protein ubiquitination sites. *Proteins* **2010**, *78*, 365–380.

22. Edwards, Y.J.; Lobley, A.E.; Pentony, M.M.; Jones, D.T. Insights into the regulation of intrinsically disordered proteins in the human proteome by analyzing sequence and gene expression data. *Genome Biol.* **2009**, *10*, R50.

23. Peng, Z.; Mizianty, M.J.; Xue, B.; Kurgan, L.; Uversky, V.N. More than just tails: Intrinsic disorder in histone proteins. *Mol. Biosyst.* **2012**, *8*, 1886–1901.

24. Xue, B.; Jeffers, V.; Sullivan, W.J.; Uversky, V.N. Protein intrinsic disorder in the acetylome of intracellular and extracellular toxoplasma gondii. *Mol. Biosyst.* **2013**, *9*, 645–657.

25. Horikoshi, M. Histone acetylation: From code to web and router via intrinsically disordered regions. *Curr. Pharm. Des.* **2013**, *19*, 5019–5042.

26. Gough, J.; Dunker, A.K. Sequences and topology: Disorder, modularity, and post/pre translation modification. *Curr. Opin. Struct. Biol.* **2013**, *23*, 417–419.

27. Huang, Q.; Chang, J.; Cheung, M.K.; Nong, W.; Li, L.; Lee, M.T.; Kwan, H.S. Human proteins with target sites of multiple post-translational modification types are more prone to be involved in disease. *J. Proteome Res.* **2014**, *13*, 2735–2748.

28. Pejaver, V.; Hsu, W.L.; Xin, F.; Dunker, A.K.; Uversky, V.N.; Radivojac, P. The structural and functional signatures of proteins that undergo multiple events of post-translational modification. *Protein Sci.* **2014**, *23*, 1077–1093.

29. Buljan, M.; Chalancon, G.; Dunker, A.K.; Bateman, A.; Balaji, S.; Fuxreiter, M.; Babu, M.M. Alternative splicing of intrinsically disordered regions and rewiring of protein interactions. *Curr. Opin. Struct. Biol.* **2013**, *23*, 443–450.

30. Romero, P.R.; Zaidi, S.; Fang, Y.Y.; Uversky, V.N.; Radivojac, P.; Oldfield, C.J.; Cortese, M.S.; Sickmeier, M.; LeGall, T.; Obradovic, Z.; *et al.* Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 8390–8395.

31. Pentony, M.M.; Jones, D.T. Modularity of intrinsic disorder in the human proteome. *Proteins* **2010**, *78*, 212–221.

32. Trudeau, T.; Nassar, R.; Cumberworth, A.; Wong, E.T.; Woollard, G.; Gsponer, J. Structure and intrinsic disorder in protein autoinhibition. *Structure* **2013**, *21*, 332–341.

33. Wang, J.; Wang, Y.; Chu, X.; Hagen, S.J.; Han, W.; Wang, E. Multi-scaled explorations of binding-induced folding of intrinsically disordered protein inhibitor ia3 to its target enzyme. *PLoS Comput. Biol.* **2011**, *7*, e1001118.

34. Uversky, V.N.; Shah, S.P.; Gritsyna, Y.; Hitchcock-DeGregori, S.E.; Kostyukova, A.S. Systematic analysis of tropomodulin/tropomyosin interactions uncovers fine-tuned binding specificity of intrinsically disordered proteins. *J. Mol. Recognit.* **2011**, *24*, 647–655.

35. Vuzman, D.; Levy, Y. Intrinsically disordered regions as affinity tuners in protein-DNA interactions. *Mol. Biosyst.* **2012**, *8*, 47–57.

36. Mittag, T.; Kay, L.E.; Forman-Kay, J.D. Protein dynamics and conformational disorder in molecular recognition. *J. Mol. Recognit.* **2010**, *23*, 105–116.

37. Bergantino, F.; Guariniello, S.; Raucci, R.; Colonna, G.; de Luca, A.; Normanno, N.; Costantini, S. Structure-fluctuation-function relationships of seven pro-angiogenic isoforms of vegfa, important mediators of tumorigenesis. *Biochim. Biophys. Acta* **2015**, *1854*, 410–425.

38. Mileo, E.; Lorenzi, M.; Erales, J.; Lignon, S.; Puppo, C.; Le Breton, N.; Etienne, E.; Marque, S.R.; Guigliarelli, B.; Gontero, B.; *et al.* Dynamics of the intrinsically disordered protein cp12 in its association with gapdh in the green alga chlamydomonas reinhardtii: A fuzzy complex. *Mol. Biosyst.* **2013**, *9*, 2869–2876.

39. Vacic, V.; Oldfield, C.J.; Mohan, A.; Radivojac, P.; Cortese, M.S.; Uversky, V.N.; Dunker, A.K. Characterization of molecular recognition features, morfs, and their binding partners. *J. Proteome Res.* **2007**, *6*, 2351–2366.

40. Oldfield, C.J.; Cheng, Y.; Cortese, M.S.; Romero, P.; Uversky, V.N.; Dunker, A.K. Coupled folding and binding with alpha-helix-forming molecular recognition elements. *Biochemistry* **2005**, *44*, 12454–12470.

41. Meszaros, B.; Simon, I.; Dosztanyi, Z. Prediction of protein binding regions in disordered proteins. *PLoS Comput. Biol.* **2009**, *5*, e1000376.

42. Disfani, F.M.; Hsu, W.L.; Mizianty, M.J.; Oldfield, C.J.; Xue, B.; Dunker, A.K.; Uversky, V.N.; Kurgan, L. Morfpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins. *Bioinformatics* **2012**, *28*, i75–i83.

43. Fang, C.; Noguchi, T.; Tominaga, D.; Yamana, H. Mfspssmpred: Identifying short disorder-to-order binding regions in disordered proteins based on contextual local evolutionary conservation. *BMC Bioinform.* **2013**, *14*, 300.

44. Malhis, N.; Gsponer, J. Computational identification of MoRFs in protein sequences. *Bioinformatics* **2015**, *31*, 1738–1744.

45. Xue, B.; Dunker, A.K.; Uversky, V.N. Retro-morfs: Identifying protein binding sites by normal and reverse alignment and intrinsic disorder prediction. *Int. J. Mol. Sci.* **2010**, *11*, 3725–3747.

46. Sun, X.; Xue, B.; Jones, W.T.; Rikkerink, E.; Dunker, A.K.; Uversky, V.N. A functionally required unfoldome from the plant kingdom: Intrinsically disordered n-terminal domains of gras proteins are involved in molecular recognition during plant development. *Plant Mol. Biol.* **2011**, *77*, 205–223.

47. Xue, B.; Oldfield, C.J.; Van, Y.Y.; Dunker, A.K.; Uversky, V.N. Protein intrinsic disorder and induced pluripotent stem cells. *Mol. Biosyst.* **2012**, *8*, 134–150.

48. Brunquell, J.; Yuan, J.; Erwin, A.; Westerheide, S.D.; Xue, B. Dbc1/ccar2 and ccar1 are largely disordered proteins that have evolved from one common ancestor. *Biomed. Res. Int.* **2014**, *2014*, 418458.

49. Lise, S.; Jones, D.T. Sequence patterns associated with disordered regions in proteins. *Proteins* **2005**, *58*, 144–150.

50. Fernandes, C.L.; Escouto, G.B.; Verli, H. Structural glycobiology of heparinase II from pedobacter heparinus. *J. Biomol. Struct. Dyn.* **2014**, *32*, 1092–1102.

51. Moroz, N.A.; Novak, S.M.; Azevedo, R.; Colpan, M.; Uversky, V.N.; Gregorio, C.C.; Kostyukova, A.S. Alteration of tropomyosin-binding properties of tropomodulin-1 affects its capping ability and localization in skeletal myocytes. *J. Biol. Chem.* **2013**, *288*, 4899–4907.

52. Wood, M.; Rae, G.M.; Wu, R.M.; Walton, E.F.; Xue, B.; Hellens, R.P.; Uversky, V.N. Actinidia DRM1—An intrinsically disordered protein whose mrna expression is inversely correlated with spring budbreak in kiwifruit. *PLoS ONE* **2013**, *8*, e57354.

53. Sun, X.; Greenwood, D.R.; Templeton, M.D.; Libich, D.S.; McGhie, T.K.; Xue, B.; Yoon, M.; Cui, W.; Kirk, C.A.; Jones, W.T.; *et al.* The intrinsically disordered structural platform of the plant defence hub protein rpm1-interacting protein 4 provides insights into its mode of action in the host-pathogen interface and evolution of the nitrate-induced domain protein family. *FEBS J.* **2014**, *281*, 3955–3979.

54. Redwan, E.M.; Xue, B.; Almehdar, H.A.; Uversky, V.N. Disorder in milk proteins: Caseins, intrinsically disordered colloids. *Curr. Protein Pept. Sci.* **2015**, *16*, 228–242.

55. Almehdar, H.A.; El-Fakharany, E.M.; Uversky, V.N.; Redwan, E.M. Disorder in milk proteins: Structure, functional disorder, and biocidal potentials of lactoperoxidase. *Curr. Protein Pept. Sci.* **2015**, *16*, 352–365.

56. Albar, A.H.; Almehdar, H.A.; Uversky, V.N.; Redwan, E.M. Structural heterogeneity and multifunctionality of lactoferrin. *Curr. Protein Pept. Sci.* **2014**, *15*, 778–797.

57. Uversky, V.N. Proteins without unique 3D structures: Biotechnological applications of intrinsically unstable/disordered proteins. *Biotechnol. J.* **2015**, *10*, 256–366.

58. Vintsyuk, T.K. Speech discrimination by dynamic programming. *Kibernetika* **1968**, *4*, 81–88.

59. Needleman, S.B.; Wunsch, C.D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **1970**, *48*, 443–453.

60. Wagner, R.A.; Fischer, M.J. The string-to-string correction problem. *J. ACM* **1974**, *21*, 168–173.

61. Anderson, C.W.; Appella, E. Signaling to the p53 tumor suppressor through pathways activated by genotoxic and nongenotoxic stress. In *Handbook of Cell Signaling*, Bradshaw, R.A., Dennis, E.A., Eds.; Academic Press: New York, NY, USA, 2004; pp. 237–247.

62. Joerger, A.C.; Fersht, A.R. Structural biology of the tumor suppressor p53. *Annu. Rev. Biochem.* **2008**, *77*, 557–582.

63. Collavin, L.; Lunardi, A.; del Sal, G. P53-family proteins and their regulators: Hubs and spokes in tumor suppression. *Cell Death Differ.* **2010**, *17*, 901–911.

64. Armstrong, J.F.; Kaufman, M.H.; Harrison, D.J.; Clarke, A.R. High-frequency developmental abnormalities in p53-deficient mice. *Curr. Biol.* **1995**, *5*, 931–936.

65. Hollstein, M.; Sidransky, D.; Vogelstein, B.; Harris, C.C. P53 mutations in human cancers. *Science* **1991**, *253*, 49–53.

66. Uversky, V.N.; Oldfield, C.J.; Midic, U.; Xie, H.; Xue, B.; Vucetic, S.; Iakoucheva, L.M.; Obradovic, Z.; Dunker, A.K. Unfoldomics of human diseases: Linking protein intrinsic disorder with diseases. *BMC Genomics* **2009**, *10* (Suppl. 1), S7.

67. Xue, B.; Brown, C.J.; Dunker, A.K.; Uversky, V.N. Intrinsically disordered regions of p53 family are highly diversified in evolution. *Biochim. Biophys. Acta* **2013**, *1834*, 725–738.

68. Uversky, A.V.; Xue, B.; Peng, Z.; Kurgan, L.; Uversky, V.N. On the intrinsic disorder status of the major players in programmed cell death pathways. *F1000Research* **2013**, *2*, 190.

69. Peng, Z.; Xue, B.; Kurgan, L.; Uversky, V.N. Resilience of death: Intrinsic disorder in proteins involved in the programmed cell death. *Cell Death Differ.* **2013**, *20*, 1257–1267.

70. Oldfield, C.J.; Meng, J.; Yang, J.Y.; Yang, M.Q.; Uversky, V.N.; Dunker, A.K. Flexible nets: Disorder and induced fit in the associations of p53 and 14–3-3 with their partners. *BMC Genomics* **2008**, *9* (Suppl. 1), S1.

71. Oates, M.E.; Romero, P.; Ishida, T.; Ghalwash, M.; Mizianty, M.J.; Xue, B.; Dosztanyi, Z.; Uversky, V.N.; Obradovic, Z.; Kurgan, L.; *et al.* D^2p^2: Database of disordered protein predictions. *Nucleic Acids Res.* **2013**, *41*, D508–D516.

72. Dosztanyi, Z.; Csizmok, V.; Tompa, P.; Simon, I. Iupred: Web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* **2005**, *21*, 3433–3434.

73. Obradovic, Z.; Peng, K.; Vucetic, S.; Radivojac, P.; Dunker, A.K. Exploiting heterogeneous sequence properties improves prediction of protein disorder. *Proteins* **2005**, *61* (Suppl. 7), 176–182.

74. Peng, K.; Radivojac, P.; Vucetic, S.; Dunker, A.K.; Obradovic, Z. Length-dependent prediction of protein intrinsic disorder. *BMC Bioinform.* **2006**, *7*, 208.

75. Ishida, T.; Kinoshita, K. Prdos: Prediction of disordered protein regions from amino acid sequence. *Nucleic Acids Res.* **2007**, *35*, W460–W464.

76. Walsh, I.; Martin, A.J.; di Domenico, T.; Tosatto, S.C. Espritz: Accurate and fast prediction of protein disorder. *Bioinformatics* **2012**, *28*, 503–509.

77. Iakoucheva, L.M.; Radivojac, P.; Brown, C.J.; O'Connor, T.R.; Sikes, J.G.; Obradovic, Z.; Dunker, A.K. The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res.* **2004**, *32*, 1037–1049.

78. Osada, M.; Ohba, M.; Kawahara, C.; Ishioka, C.; Kanamaru, R.; Katoh, I.; Ikawa, Y.; Nimura, Y.; Nakagawara, A.; Obinata, M.; *et al.* Cloning and functional analysis of human p51, which structurally and functionally resembles p53. *Nat. Med.* **1998**, *4*, 839–843.

79. Yang, A.; Kaghad, M.; Wang, Y.; Gillett, E.; Fleming, M.D.; Dotsch, V.; Andrews, N.C.; Caput, D.; McKeon, F. P63, a p53 homolog at 3q27–29, encodes multiple products with transactivating, death-inducing, and dominant-negative activities. *Mol. Cell* **1998**, *2*, 305–316.

80. Kaghad, M.; Bonnet, H.; Yang, A.; Creancier, L.; Biscan, J.C.; Valent, A.; Minty, A.; Chalon, P.; Lelias, J.M.; Dumont, X.; *et al.* Monoallelically expressed gene related to p53 at 1p36, a region frequently deleted in neuroblastoma and other human cancers. *Cell* **1997**, *90*, 809–819.

81. Yang, A.; Kaghad, M.; Caput, D.; McKeon, F. On the shoulders of giants: P63, p73 and the rise of p53. *Trends Genet.* **2002**, *18*, 90–95.

82. Murray-Zmijewski, F.; Lane, D.P.; Bourdon, J.C. P53/p63/p73 isoforms: An orchestra of isoforms to harmonise cell differentiation and response to stress. *Cell Death Differ.* **2006**, *13*, 962–972.

83. Xue, B.; Dunbrack, R.L.; Williams, R.W.; Dunker, A.K.; Uversky, V.N. Pondr-fit: A meta-predictor of intrinsically disordered amino acids. *Biochim. Biophys. Acta* **2010**, *1804*, 996–1010.

84. He, B.; Wang, K.; Liu, Y.; Xue, B.; Uversky, V.N.; Dunker, A.K. Predicting intrinsic disorder in proteins: An overview. *Cell Res.* **2009**, *19*, 929–949.

85. Sickmeier, M.; Hamilton, J.A.; LeGall, T.; Vacic, V.; Cortese, M.S.; Tantos, A.; Szabo, B.; Tompa, P.; Chen, J.; Uversky, V.N.; *et al.* Disprot: The database of disordered proteins. *Nucleic Acids Res.* **2007**, *35*, D786–D793.

86. Myers, C.S.; Rabiner, L.R. A comparative-study of several dynamic time-warping algorithms for connected-word recognition. *Bell Syst. Tech. J.* **1981**, *60*, 1389–1409.