

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/376409461>

IDPpred: a new sequence-based predictor for identification of intrinsically disordered protein with enhanced accuracy

Article in Journal of Biomolecular Structure and Dynamics · December 2023

DOI: 10.1080/07391102.2023.2290615

CITATION

1

READS

33

5 authors, including:



Rajkrishna Mondal
Nagaland University

18 PUBLICATIONS 125 CITATIONS

SEE PROFILE



Tapobrata Lahiri
Indian Institute of Information Technology Allahabad

59 PUBLICATIONS 233 CITATIONS

SEE PROFILE



Asmita Tripathi
Indian Institute of Information Technology Allahabad

2 PUBLICATIONS 3 CITATIONS

SEE PROFILE

IDPpred: a new sequence-based predictor for identification of intrinsically disordered protein with enhanced accuracy

Deepak Chaurasiya, Rajkrishna Mondal, Tapobrata Lahiri, Asmita Tripathi & Tejas Ghinmine

To cite this article: Deepak Chaurasiya, Rajkrishna Mondal, Tapobrata Lahiri, Asmita Tripathi & Tejas Ghinmine (11 Dec 2023): IDPpred: a new sequence-based predictor for identification of intrinsically disordered protein with enhanced accuracy, Journal of Biomolecular Structure and Dynamics, DOI: [10.1080/07391102.2023.2290615](https://doi.org/10.1080/07391102.2023.2290615)

To link to this article: <https://doi.org/10.1080/07391102.2023.2290615>



View supplementary material [↗](#)



Published online: 11 Dec 2023.



Submit your article to this journal [↗](#)



View related articles [↗](#)



View Crossmark data [↗](#)



IDPpred: a new sequence-based predictor for identification of intrinsically disordered protein with enhanced accuracy

Deepak Chaurasiya^a, Rajkrishna Mondal^b, Tapobrata Lahiri^a, Asmita Tripathi^a and Tejas Ghinmine^a

^aDepartment of Applied Sciences, Indian Institute of Information Technology, Prayagraj, UP, India; ^bDepartment of Biotechnology, Nagaland University, Dimapur, Nagaland, India

Communicated by Ramaswamy H. Sarma

ABSTRACT

Discovery of intrinsically disordered proteins (IDPs) and protein hybrids that contain both intrinsically disordered protein regions (IDPRs) along with ordered regions has changed the sequence–structure–function paradigm of protein. These proteins with lack of persistently fixed structure are often found in all organisms and play vital roles in various biological processes. Some of them are considered as potential drug targets due to their overrepresentation in pathophysiological processes. The major bottlenecks for characterizing such proteins are their occasional overexpression, difficulty in getting purified homogeneous form and the challenge of investigating them experimentally. Sequence-based prediction of intrinsic disorder remains a useful strategy especially for many large-scale proteomic investigations. However, worst accuracy still occurs for short disordered regions with less than ten residues, for the residues close to order–disorder boundaries, for regions that undergo coupled folding and binding in presence of partner, and for prediction of fully disordered proteins. Annotation of fully disordered proteins mostly relies on the far-UV circular dichroism experiment which gives overall secondary structure composition without residue-level resolution. Current methods including that using secondary structure information failed to predict half of target IDPs correctly in the recent Critical Assessment of protein Intrinsic Disorder prediction (CAID) experiment. This study utilized profiles of random sequential appearance of physicochemical properties of amino acids and random sequential appearance of order and disorder promoting amino acids in protein together with the existing CIDER feature for the prediction of IDP from sequence input. Our method was found to significantly outperform the existing predictors across different datasets.

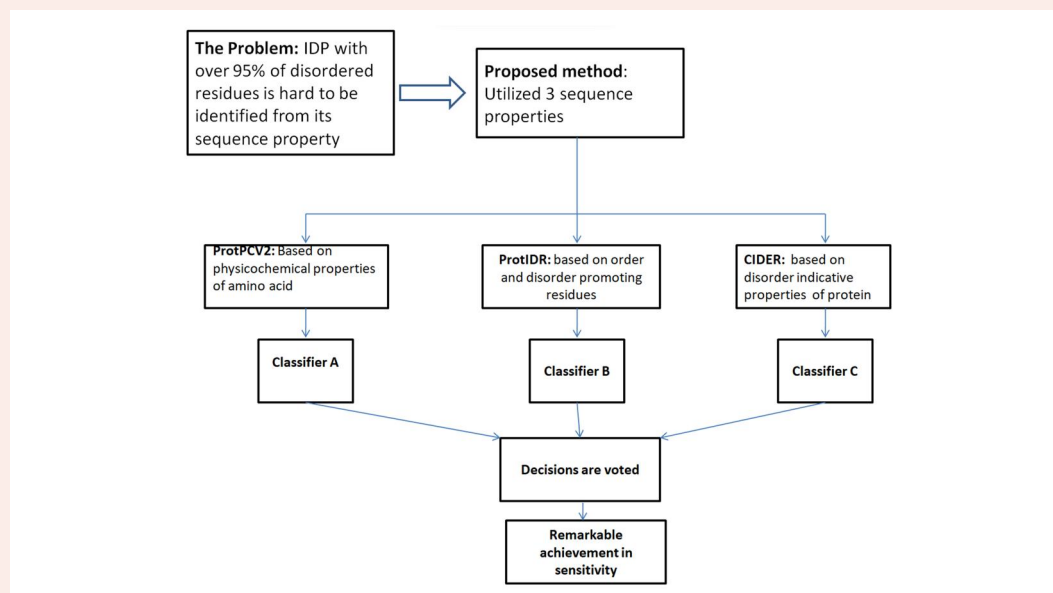
ARTICLE HISTORY

Received 27 July 2023

Accepted 15 November 2023

KEYWORDS

Intrinsically disordered protein; numerical representation of sequence; periodicity count value and predictor



1. Introduction

Under physiological conditions some proteins or their regions adopt variable conformations which is considered as an exception of Anfinsen dogma (Anfinsen, 1973). Abundance of glycine and proline residues, low overall hydrophobicity and large net charge are the main causes of disorder in a protein (Cheng et al., 2010; Uversky, 2019). However, the structural ambiguity of the proteins is context dependent (Mohan et al., 2009). Based on the richness of disordered residues, proteins can be broadly classified into three variants. First example of proteins has negligible disorder that adopts a fixed, three-dimensional fold. The second variant comprises of structured proteins harboring intrinsically disordered regions (IDPRs) with less than 30% of disordered residues in their sequence that are over-represented among enzymes, transporters, cell-junction, cell adhesion, receptors, enzyme modulators, cytoskeletal proteins. The third type refers to intrinsically disordered proteins (IDPs) having more than 30% of disordered residues which show some distinct functions such as nucleic acid binding proteins, chromatin binding proteins, transcription factors, and developmental processes (Deiana et al., 2019). However, according to the recent finding under CAID2018 competition (Critical Assessment of Protein Intrinsic Disorder Prediction) a protein containing at least 95% of disordered residues is considered as IDPs (Necci et al., 2021). IDPs are widely distributed across life forms such as eukaryotes, prokaryotes and viruses and perform multiple biological functions. In last two decades various reports corroborates their significant involvement in human diseases such as genetic diseases (Midic et al., 2009), cancers (Iakoucheva et al., 2002), cardiovascular diseases (Cheng et al., 2006), neurodegenerative disorder (Raychaudhuri et al., 2009), Alzheimer's disease (Uversky et al., 2009) and many more diseases (Uversky et al., 2008). Some of them are considered as promising targets for drug discovery (Metallo, 2010). Currently, like ordered proteins, experimental techniques have been used to decipher structural state of Intrinsically disordered proteins, for examples, NMR (Nuclear Magnetic Resonance), X-Ray crystallography, CD (circular dichroism) spectroscopy, SAXS (small-angle X-ray scattering), smFRET (single molecule Fluorescence Resonance energy transfer) (Receveur-Bréchet et al., 2006; van der Lee et al., 2014). However, the issues like, overexpression of these proteins, problems in their purification and challenges to investigate them experimentally are hindering structural and functional characterization of such proteins. Sequence-based protein disorder prediction remains a useful strategy especially for high throughput proteomic investigations. There exists many protein disorder prediction methods based on different principles and computing techniques e.g., the methods relying on statistics of physicochemical properties of the amino acids, secondary structure information and template based information. Almost all of these works utilized standard machine learning approaches while in some cases meta predictors were also utilized by assimilating multiple classifiers (Deng et al., 2012; Liu et al., 2019). From the survey of existing models for the prediction of protein disorder, there appears to be two distinct observations. First one is the CASP and CAID coordinated activities to get better prediction models with enriched functional pieces of information and the second one is based on

development of meta predictors combining multiple predictors for further enhancement of the accuracy. Prediction of fully disordered proteins is considered as one of the most challenging categories as set by CAID, 2018 since, it is difficult to investigate residue-level information of disorder through experimental methods (Necci et al., 2021). Towards this task, the frequently used circular dichroism spectroscopy evaluates overall secondary structure, folding and binding properties of proteins. However, it is not found to be capable of providing individual residue level information about disorder. Current methods including that using secondary structure information failed to predict half of the target IDPs correctly in the recent Critical Assessment of protein Intrinsic Disorder prediction (CAID) competition probably due to lack of further analysis of the sequential location based study of these secondary structures. Furthermore, in-depth analysis on roles of multiple disordered residues vis-à-vis the structural stability of a protein appears to be pending yet. Stemmed from these above-given facts, it appears that there is room for improvement in the disorder prediction methods to enhance accuracy significantly.

In this context, this work utilized the locational profile of physicochemical properties within a protein sequence, ProtPCV2 following the work of Tripathi et al. (2023) to build an improved classifier. Similarly, a classifier was built using novel locational profile of order and disorder promoting amino acids (ProtIDR). Also, one more classifier was employed that utilized existing CIDER representation (Holehouse et al., 2017) as features. Finally, the output obtained through voting using three outputs of these predictors was utilized for final prediction of IDP.

2. Materials and methods

2.1. Programming and scripting language used for data processing

All the proteins data for this work were collected from the CAID2018 competition (Critical Assessment of Protein Intrinsic Disorder Prediction) a protein containing at least 95% of disordered residues are considered as IDPs (Necci et al., 2021). The details were however given in Section 2.3. For the processing of these datasets, the scripts and routines were written under the MATLAB programming environment of version: '9.7.0.1586710 (R2019b) Update 8'.

2.2. Sequence based feature extraction

2.2.1. ProtIDR based feature extraction for identification of IDPs

The novel ProtIDR feature was extracted following the algorithm similar to that demonstrated by Pal et al. (2020) and Tripathi et al. (2023) for computation of the ProtPCV2 feature. Towards this direction, we used mostly cited

Table 1. Amino-acids classification based on disordered promotion.

Pcp	Nature	Amino acids	Location marks in the sequence
DPaa		A,R,G,Q,S,P,E,K	1
OPaa		N,W,C,I,F,Y,V,L,M,D,H,T	0

classification information about amino acids in the context of IDPs which are as follows:

- i. Order promoting amino acids (OPaa) and,
- ii. disorder promoting amino acids (DPaa).

The algorithmic steps for the calculation of inter-arrival distances (IADs) for successive DPaa residues are explained below using Table 1.

Step 1: For, a protein sequence, $S = \{a_1, a_2, a_3, a_4, \dots, a_N\}$, where a_1, a_2, \dots are the successive residues, we get the binary sequence, $f(S)$ through an Indicator function, f , as,

$$f(S) = \{f(a_1), f(a_2), f(a_3), f(a_4), \dots, f(a_N)\}$$

where, $f(a_j) = 1$ if $a_j \in \text{DPaa}$ else 0 if $a_j \in \text{OPaa}$

Step 2: The rule for computation of IAD values for the '1' elements from the binary sequence, $f(S)$ is demonstrated from the following example:

1. For '11', IAD is 0
2. For '101', IAD is 1,
3. For '1001', IAD is 2, and so on.

The successive IAD values generated for '1' elements of $f(S)$ through this method builds the array, IAD^1 . Similarly, the successive IAD values generated for '0' elements of $f(S)$ gives the array IAD^0 .

The steps for building each of the IAD arrays were as follows:

Step 1: First a frequency histogram of IADs was built choosing a range of IAD values from 0 to 14 since the maximally probable inter-arrival distance was found as 14.

Step 2: The histogram referred in Step 1 was made with heuristically chosen five intervals. This heuristics followed from the result of accuracies compared with those obtained using different numbers of intervals.

Step 3: The frequency histogram obtained in Step 1 was converted to probability distribution using standard statistical procedure.

Thus for the above-mentioned two IAD arrays, we get 10 probability values which constitutes, ProtIDR array.

2.2.2. ProtPCV2 based feature extraction from a protein sequence for identification of IDPs

Similarly, the ProtPCV2 feature was computed for the five physicochemical properties of a protein to yield five binary

arrays using the criteria as shown in the following Table 2 following the works of Pal et al. (2020) and Tripathi et al. (2023):

It is evident that the five binary arrays thus yielded by using these five physicochemical properties will produce an array of 50 probability values to be referred as ProtPCV2 feature as demonstrated in section 2.2.1.

2.2.3. CIDER based feature extraction for identification of IDPs

CIDER is a web-server which provides various disorder indicative properties of a protein based on their sequences (Holehouse et al., 2017). The 10 parameters extracted for this work using CIDER were f^- (fraction of negative charge) and f^+ (fraction of positive charge) value from Das-Pappu phase diagram, FCR (fraction of charged residues), NCPR (net charge per residue), Kappa (κ), Omega (Ω), Sigma (σ), Delta (δ), Max Delta, and Hydropathy sequence features. CIDER_Kappa parameters describe the extent of charged amino acid mixing in a sequence. The patterning parameter Omega (Ω) is analogous to the parameter Kappa (κ). It actually describes the patterning of charged + proline residues in a protein sequence with respect to all other residues. The Hydropathy value referred to here is the Kyte-Doolittle hydropathy value that is the residue average value for the entire sequence within a range from 0 (least hydrophobic) to 9 (most hydrophobic).

2.3. Datasets

The limitation of previously developed IDPs predictors was the availability of less experimental data (less experimentally confirmed IDPs). A training dataset consists of 142 unique fully disordered proteins with over 95% of disordered residues were collected from Disprot 22_12 release (<https://disprot.org/>). Also, experimentally annotated 300 control data for proteins (comprising of proteins which are stable and remaining disordered proteins which are not fully disorder) were collected from DisProt and MobiDB (meta-predictor). The control data comprises of the proteins of various organisms such as human, mouse, rat, zebra fish, and fruit fly. Furthermore, for these control data, disorderedness of the proteins were also validated and later selected using freely available and mostly cited disordered predictor methods such as PONDR pool (PONDR FIT, PONDR VLS2, PONDR VLXT), ESpritz (NMR, X-Ray, and Disprot). The selection of disordered proteins was done by the nomenclature suggested by Deiana et al. (2019). After converting these proteins to any one of their corresponding feature vectors (ProtPCV2, ProtIDR or CIDER), we divided these proteins at random in a 70 by 30 ratio into two disjoint datasets: the training dataset (70%) to train the machine learning models and test dataset (30%) to compare the prediction performance. For the purpose of testing, the CAID dataset of 652 proteins (Necci et al., 2021) comprising of 45 fully disordered proteins and remaining control proteins (comprising of stable and remaining disordered proteins which are not fully disorder) were

Table 2. Amino-acids classification based on physicochemical properties.

Pcp	Sub-property	Amino acids	
HC	Aromatic(A)	F,W,Y,H	1
	Aliphatic(AP)	A,M,C,L,V,I,D,E,N,Q,S,T,R,K,P,G	0
PO	Polar(P)	D,E,H,K,N,Q,R,S,T,Y	1
	Non-Polar(NP)	A,C,F,G,I,L,M,P,W,V	0
CC	Non-Zero(NZ)	C,I,D,E,N,Q,Y,S,T,R,K,H,P,W,G	1
	Zero(Z)	A,F,L,M,V	0
HI	Hydrophobic(HB)	A,M,C,F,L,V,I	1
	Hydrophilic(HP)	D,E,N,Q,Y,S,T,R,K,H,P,W,G	0
HY	Positive(PO)	A,C,F,G,I,P,W,S,T,L,V	1
	Negative(NE)	D,E,Q,Y,R,K,H,M,N	0

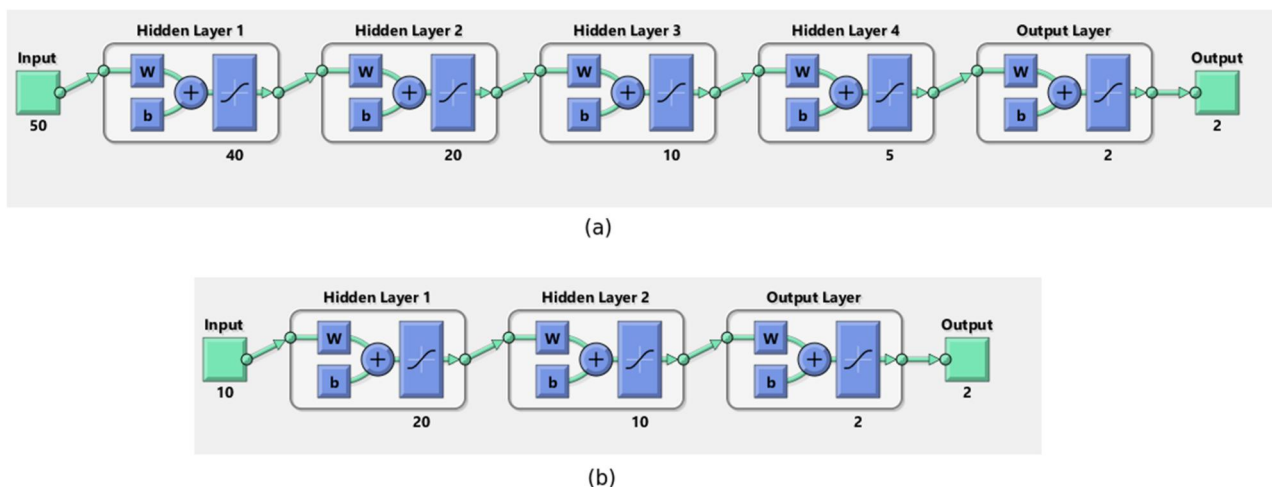


Figure 1. BPN architecture that uses (a) 50 dimensional ProtPCV and (b) 10 dimensional ProtIDR or CIDER features.

utilized. It should also be pointed out that CAID data were drawn from DisProt data. Therefore, only the remaining DisProt data were utilized for training purposes along with the above-mentioned pool of training data.

2.4. Design and implementation of IDPpred using three classifiers for identification of fully disordered proteins

In this work, emphasis was given to find features which are well descriptive of disorderedness of a protein having reasonably high discriminating power between the following two target classes:

- i. Fully disordered proteins with over 95% of disordered residues with class label F_disord.
- ii. Control data for proteins (comprising of proteins which are stable and remaining disordered proteins which are not fully disorder) with class label Control.

First three separate classifiers were designed using back-propagation network (BPN) protocol with the help of the training datasets as described in Section 2.3 taking following different features as inputs to these three different classifiers:

- i. ProtPCV2
- ii. ProtIDR
- iii. CIDER

The motivation was to employ physicochemically meaningful features that have substantial potential to capture disorderedness present within the structure of a protein so that even a simple classifier can discriminate between fully disordered and control set of proteins. As described in the Figure 1, the following network architectures were recruited for the intended classification tasks:

These three classifiers were finally used separately to classify the test datasets on the basis of the three features as described earlier. Finally to enhance the classification accuracy further the outputs collected from each of these

classifiers were pressed for voting with the following voting principle to build the IDPpred predictor:

- For class label F_disord, output is designated as '1'
- For class label Control, output is designated as '0'
- If sum of the outputs generated by 3 classifiers ≥ 2 , the class label for the test data is F_disord, else, Control

2.5. Accuracy parameters used to evaluate the predictive performance of IDPpred and other predictors

For the assessment of predictive performance, we used following measures that are similar to that used in CASP8 and previous versions of CASPs (Bordoli et al., 2007; Jin & Dunbrack, 2005; Melamud & Moul, 2003; Meng et al., 2017; Monastyrskyy et al., 2014; Noivirt-Brik et al., 2009; Walsh et al., 2015):

1. Sensitivity ($\text{Sensitivity} = TP/N_{\text{dis}}$),
2. Accuracy [$\text{Accuracy} = (\text{Sensitivity} + \text{Specificity})/2$],
3. Selectivity ($\text{Selectivity} = TP/TP + FP$),
4. Specificity ($\text{Specificity} = TN/TN + FN$),
5. F-measure [$F = 2 \times \text{Sensitivity} \times \text{Selectivity} / (\text{Sensitivity} + \text{Selectivity})$],
6. Matthews correlation coefficient (MCC), and
7. Score ($S_w = \text{Sensitivity} + \text{Specificity} - 1$)

where, TP, TN, FN, and FP are the number of true positives, true negatives, false negatives, and false positives respectively.

3. Results

3.1. ProtPCV2 vector for a representative target protein

Each of the protein sequences was converted into 50 valued fixed dimensional numerical representation (ProtPCV2) as described in Tripathi et al. (2023). Conversion on protein sequence to ProtPCV2 value was given below for a CAID target protein, DP01295:

Sequence:

MEPVDPKLEPWKHPGSPKTAACNNCYCKRCCLHCQVCFMTKGLGI-
YYGRKKRRRRRASQDRQTHQDSLSEQ

Corresponding ProtPCV2 array:

{92.85,7.14,0,0,0,92.85,0,0,7.14,0,33.33,16.66,33.33,0,16.66,100,
0,0,0,0,72.22,11.11,0,5.55,11.11,96.15,3.84,0,0,0,100,0,0,0,30,
30,0,20,20,90.90,6.06,0,3.03,0,83.78,16.22,0,0,0}

3.2. ProtIDR vector for a representative target protein

Every protein sequence was converted into 10 valued fixed dimensional numerical representation (ProtIDR) as mentioned in the materials and methods Section 2.1.2 ProtIDR value for the DP01295 was cited below:

Corresponding ProtIDR array:

{93.55,0,3.23,0,3.23,89.74,7.69,2.56,0,0}

3.3. CIDER vector for a representative target protein

Similarly a 10 valued features for the query protein sequences were extracted using CIDER web server based output (Holehouse et al., 2017) and result against DP01295 given below:

Corresponding CIDER array:

{72,0.08,0.21,0.29,0.13,0.38,0.25,0.05,0.09,0.23,3.26}

3.4. Prediction of intrinsic disorder

Fully disordered proteins have unique biological functions such as nucleic acid binding, chromatin binding, serving as transcription factors, and regulation of developmental processes. Moreover, they are very difficult to be probed through experimental methods up to atomic level resolution. To accomplish the task to differentiate fully disordered proteins from control proteins of the CAID dataset we have taken nonconventional location level information of individual residues rather than gross statistics of the whole set of residues considered by other predictors. Outputted result from our method was compared with the prediction by fIDPNN which uses deep feedforward neural network to predict protein disorder from sequence information profile into three-level hierarchy: residue-level, window-level and protein-level based on relevant structural and functional information extracted utilizing popular and fast bioinformatics tools (Hu et al., 2021). Additionally, balanced accuracy (BAC) metric was used to evaluate the performance of classification models as there is an unbalanced distribution of classes in the CAID dataset. Firstly, ProtPCV, ProtIDR and CIDER features were extracted for all protein sequences as mentioned in methodology Sections 2.2.1–2.2.3. Subsequently, networks for individual predictors were trained. The three classifiers using ProtPCV2, ProtIDR and CIDER features predicted 33, 32, and 35 fully disordered proteins out of 45 CAID challenge test data with BAC 0.7216, 0.755 and 0.7935 compared to the best CAID

predictor fIDPnn that predicted 26 IDPs and secured BAC = 0.776 (Hu et al., 2021; Necci et al., 2021). Finally, a significant improvement in IDP prediction was achieved by the voting-based meta-predictor, IDPpred introduced in this work that correctly predicted 37 fully disordered proteins out of 45 CAID challenge data and achieved a BAC = 0.8353. The flowchart shown in Figure 2 depicted the workflow of IDPpred predictor while Table 3 listed the performance parameters for all the important predictors. While it is evident that, IDPpred is far ahead of the best CAID predictor fIDPnn in correctly predicting fully disordered proteins of the CAID challenge dataset, the performance of fIDPnn is slightly superior to IDPpred in predicting the control proteins. It is also noteworthy that fIDPnn predictor utilized deep learning protocol which is considered to be reasonably superior to the simple BPN network utilized in case of IDPpred. Therefore, it does not seem to be an over-expectation that the addition of a deep learning protocol in IDPpred would significantly enhance the accuracy of IDPpred predictor. It indicates the need of assimilating the complementary potential of both of these predictors for future goal. This is also demonstrated in Figures 3 and 4 showing the win of IDPpred over fIDPnn and vice versa respectively in terms of NCPR distribution, fIDPnn disorder propensity, secondary structure predicted by JPRED4 tool and superimposition of fully or partially experimentally determined structures.

4. Discussion

From the survey of the published works in the last two decades on intrinsically disordered proteins, there appear to be two distinct observations. The first one is the community-based effort to develop and implement evaluation strategies of protein disorder prediction coordinated by CASP 5th to 10th editions (Bordoli et al., 2007; Jin & Dunbrack, 2005; Melamud & Moul, 2003; Monastyrskyy et al., 2014; Noivirt-Brik et al., 2009) and followed by more specialized CAID (Necci et al., 2021). The second observation is that there exists plenty of room for improvement in the prediction strategies as most of the current methods failed to predict even half of the targets given in recent CAID (Necci et al., 2021). Distinct functions of the proteins with disorder regions and fully disordered proteins are increasingly apparent. The definition of the IDP proteins is changing with time. Previously proteins with overall 30% disordered residues were considered as fully disordered (Deiana et al., 2019). However in the recent CAID, proteins with disorder annotation covering at least 95% of the sequence are considered fully disordered. CAID IDP targets were highly diverse with respect to their length (24 to 699 aa) and physicochemical properties e.g., pI (3.65 to 12.41), grand average of hydropathicity (−1.9710 to 0.3420), percentage of glycine (0 to 18.5%) and proline (0 to 14%) residues as mentioned in Supplementary Table 1. Prediction of the fully disordered proteins is considered to be most challenging because of unavailability of experimental data for validation as it is very

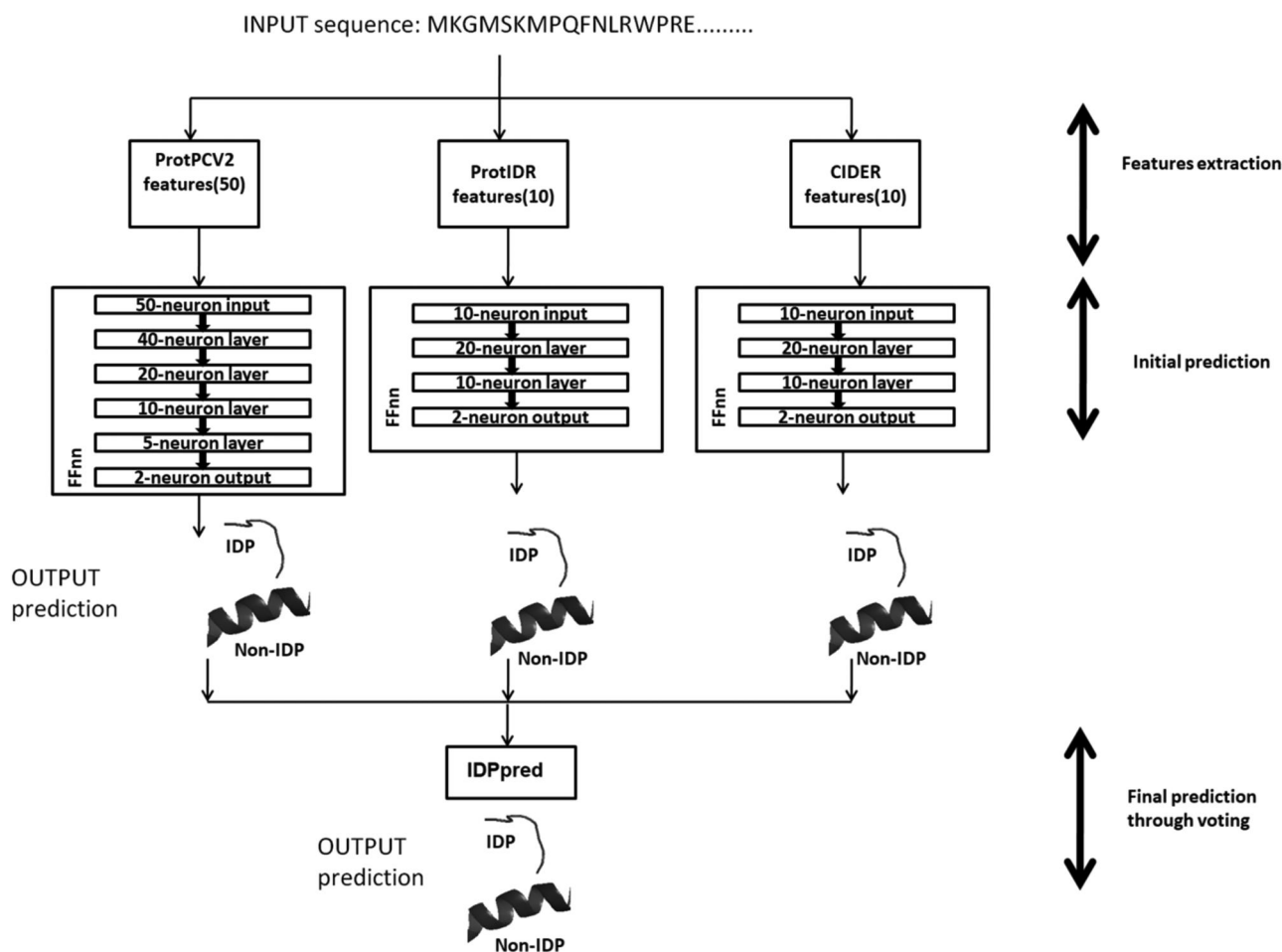


Figure 2. Flowchart depicting the workflow of IDPpred predictor.

Table 3. Result of performance of different predictors in correctly identifying fully disordered protein (disorder residues >95%) for the dataset provided by CAID.

Predictors	Performance evaluation parameters*									
	TP	TN	FP	FN	TNR	TPR	BAC	MCC	F1-s	PPV
IDPpred	37	515	92	8	0.8484	0.8222	0.8353	0.4267	0.4252	0.2868
CIDER	35	492	115	10	0.7777	0.8105	0.7935	0.3543	0.2333	0.3590
fIDPnn	26	585	16	19	0.973	0.578	0.776	0.569	0.598	0.619
ProtIDR	32	488	119	13	0.71	0.80	0.755	0.3295	0.3316	0.2119
ProtPCV2	33	431	176	12	0.7333	0.710	0.7216	0.2408	0.2598	0.157

*TN, true negatives count; TP, true positives count; FN, false negatives count; FP, false positives count; F1-s, F1-score; TNR, true negative rate, specificity; TPR, true positive rate, recall; PPV, positive predictive value, precision; BAC= (sensitivity + specificity)/2, balanced accuracy for prediction of fully disordered proteins. Proteins with disorder prediction or disorder annotation covering at least 95% of the sequence are considered fully disordered. Predictors are sorted by their BAC.

difficult to study these proteins through X-ray crystallography or NMR spectroscopy.

Prediction of IDP based on gross statistics (e.g., average) of total residue dependent characteristics seems to be oversimplified. However, the majority of the IDP are confirmed by CD experiment which gives gross protein structural information e.g., the secondary structure contents. Additionally, the role of long distance interactions on protein disorder is usually ignored. In this work, profiles of locational appearance of physicochemical properties of amino acids and profiles of locational appearance of order and disorder

promoting amino acids in protein were utilized that in general provides the spatial distribution of physicochemical properties of a protein. Since stability of a protein structure primarily depends on the distribution of amino acids carrying specific physicochemical properties within the 3D structure space, therefore, it was of prime interest to study the potential of the related features, ProtPCV2 and ProtIDR in correctly identifying a fully disordered protein. Also, because of the proven potential of existing CIDER features to offer the gross statistical signatures of a protein, it was also used for prediction of IDP from sequence input. It will not be out of the context to inform that some tools use alignment based methods to find the sequence conservation in model organisms to uplift the accuracy of disorder prediction (Ishida & Kinoshita, 2007; Yang et al., 2005). In our previous works ProtPCV and ProtPCV2 have shown higher accuracy to find better homologues compared to BLASTP and PSIBLAST (Jones & Swindells, 2002). fIDPnn, ranked as the best algorithm for the prediction of fully IDP as per recent CAID challenge that could detect 26 out of 45 targets. ProtPCV2, ProtIDR and CIDER based predictors individually have shown higher sensitivity with true positive detection of 33, 32 and 35 respectively. Accuracy was further improved by combining the outputs of these predictors through a simple voting strategy as described in Section 2.4. Analysis of the false negative target proteins revealed their atypical nature with a

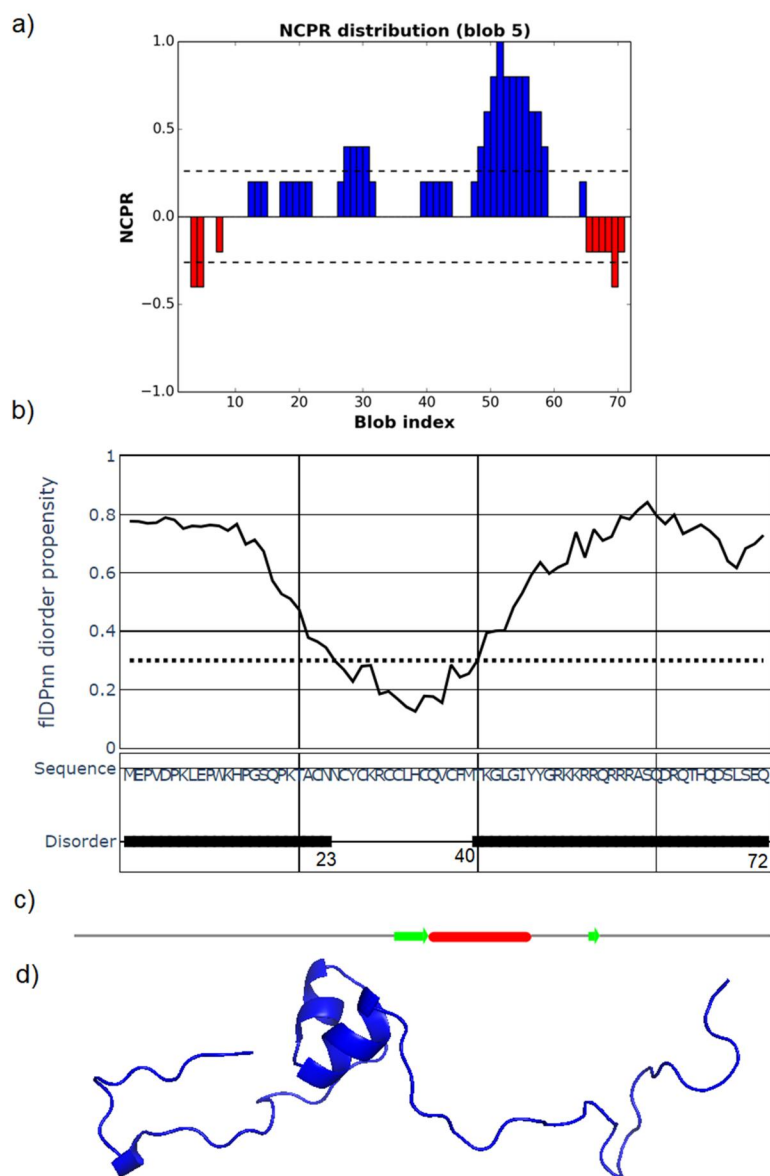


Figure 3. IDPpred wins over fIDPnn. (a) Linear net charge per residue (NCPR) diagram generated by CIDER against the query DP01295. (b) fIDPnn disorder prediction for the target unable to sense as fully disordered protein. (c) Secondary structure prediction by JPRED4 tool annotated secondary structure against query sequence (the performance was not listed in Table 3). (d) Ribbon presentation of the model structure predicted by homology modelling using modeller tool utilizing crystal structure of HIV-1 tat (PDB ID: 3MI9, chain: C).

grand average of hydropathicity (Aliphatic index and GRAVY value) more than that of the true positive proteins. Target with shorter length (24 aa) was not detected rightly by IDPpred, ProtPCV2 and ProtIDR however, the same was detected truly by CIDER based predictor. Examples of win of ProtIDR over fIDPnn and vice versa predicting fully IDP protein were represented in Figures 3 and 4, respectively which indicated a potential of complementary use of these two predictors for future goal. Secondary structures for the two CAID targets (DP01295 and DP02267) were predicted using JPRED4(<https://www.compbio.dundee.ac.uk/jpred/>) tools (Figures 3c and 4c). Detection of stable secondary structure by the JPRED4 tools proved that IDP prediction methods that rely on secondary structure information may have disadvantages for successful prediction. Probably, loss of ProtPCV2

and ProtIDR based predictor for the detection of DP02267 were because of the short size of the protein. However, the same target was successfully detected by the CIDER-based predictor.

Overall, it is evident that the performance of IDPpred predictor is competitive with the current best predictor, fIDPnn with a clear advantage of better prediction performance for the fully IDP. Since, fIDPnn utilized deep learning network, it appears that the competitive performance of IDPpred with fIDPnn is because of the potential of the features utilized by it, especially the ProtPCV2 and ProtIDR that can capture the spatial distribution of individual amino acids carrying information on specific physicochemical properties. Therefore, it does not seem to be an overexpectation that under the deep learning paradigm, IDPpred should perform much better.

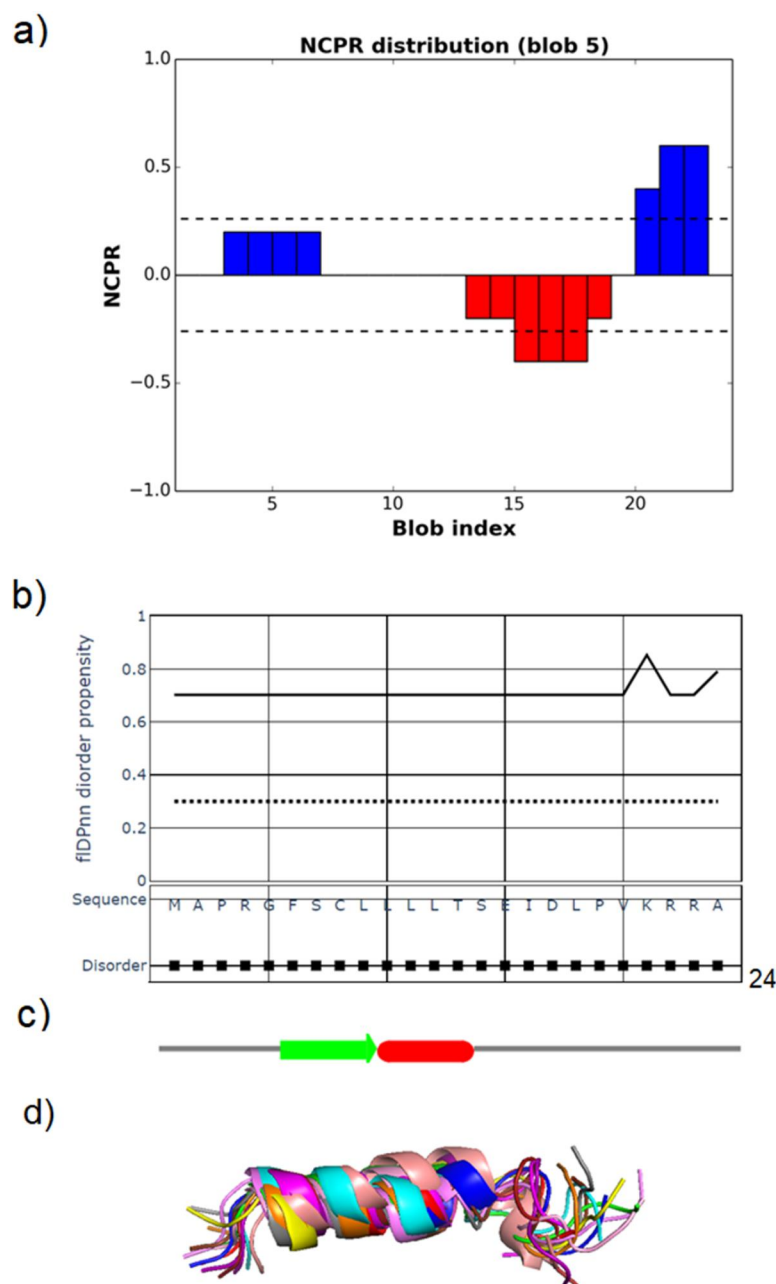


Figure 4. fIDPnn wins over IDPpred. (a) Linear net charge per residue (NCPR) diagram generated by CIDER against the query DP02267. (b) fIDPnn disorder prediction for the target successfully sensed the fully disordered protein. (c) Secondary structure prediction by JPRED4 tool annotated secondary structure against query sequence. (d) Superimposition of the Ribbon presentation of the models determined by NMR for humanin (PDB ID: 1Y32) showing structural ambiguity.

5. Conclusion

IDPs are considered as an important topic of research for its unique functional properties and involvement in pathophysiology of several life threatening human diseases. In silico based prediction method is considered as a useful strategy for high throughput IDP detection as their experimental characterization seems to be very challenging. Despite significant advances in IDP prediction from sequence information, most of the predictors participated in CAID 2018 competition failed to predict even half of the targets successfully. Our three individual methods individually and combinedly showed better performance over the best predictor in the competition for detection of fully disordered proteins against CAID targets. The significantly better performance of

the novel predictor IDPpred in this work appears to be stemmed from the potential of features ProtPCV2 and ProtIDR utilized by it that could successfully extract the information on the spatial distribution of amino acids carrying specific physicochemical properties to outperform the predictors based on gross statistics of whole residue properties of proteins.

Acknowledgment

The authors appreciatively acknowledge the resources and infrastructures provided by the Indian Institute of Information Technology Allahabad to carry out this work. Mr. Deepak Chaurasiya is grateful for the scholarship provided to him by Indian Institute of Information Technology, Allahabad.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The author(s) reported there is no funding associated with the work featured in this article.

Author contributions

Deepak Chaurasiya prepared the database, performed the experiment; Rajkrishna Mondal prepared initial draft writing and figures, Tapobrata Lahiri designed the experiments and written final draft, Asmita Tripathi and Tejas Ghinmine helped in experiment and coding.

Data availability statement

All the data used in this work are available with the corresponding author and can be shared on such demand.

References

- Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science (New York, N.Y.)*, 181(4096), 223–230. <https://doi.org/10.1126/science.181.4096.223>
- Bordoli, L., Kiefer, F., & Schwede, T. (2007). Assessment of disorder predictions in CASP7. *Proteins*, 69 Suppl 8(S8), 129–136. <https://doi.org/10.1002/prot.21671>
- Cheng, S., Cetinkaya, M., & Gräter, F. (2010). How sequence determines elasticity of disordered proteins. *Biophysical Journal*, 99(12), 3863–3869. <https://doi.org/10.1016/j.bpj.2010.10.011>
- Cheng, Y., LeGall, T., Oldfield, C. J., Dunker, A. K., & Uversky, V. N. (2006). Abundance of intrinsic disorder in protein associated with cardiovascular disease. *Biochemistry*, 45(35), 10448–10460. <https://doi.org/10.1021/bi060981d>
- Deiana, A., Forcelloni, S., Porrello, A., & Giansanti, A. (2019). Intrinsically disordered proteins and structured proteins with intrinsically disordered regions have different functional roles in the cell. *PloS One*, 14(8), e0217889. <https://doi.org/10.1371/journal.pone.0217889>
- Deng, X., Eickholt, J., & Cheng, J. (2012). A comprehensive overview of computational protein disorder prediction methods. *Molecular bioSystems*, 8(1), 114–121. <https://doi.org/10.1039/c1mb05207a>
- Holehouse, A. S., Das, R. K., Ahad, J. N., Richardson, M. O. G., & Pappu, R. V. (2017). CIDER: Resources to analyze sequence-ensemble relationships of intrinsically disordered proteins. *Biophysical Journal*, 112(1), 16–21. <https://doi.org/10.1016/j.bpj.2016.11.3200>
- Hu, G., Katuwawala, A., Wang, K., Wu, Z., Ghadermarzi, S., Gao, J., & Kurgan, L. (2021). fDPnn: Accurate intrinsic disorder prediction with putative propensities of disorder functions. *Nature Communications*, 12(1), 4438. <https://doi.org/10.1038/s41467-021-24773-7>
- Iakoucheva, L. M., Brown, C. J., Lawson, J. D., Obradović, Z., & Dunker, A. K. (2002). Intrinsic disorder in cell-signaling and cancer-associated proteins. *Journal of Molecular Biology*, 323(3), 573–584. [https://doi.org/10.1016/s0022-2836\(02\)00969-5](https://doi.org/10.1016/s0022-2836(02)00969-5)
- Ishida, T., & Kinoshita, K. (2007). PrDOS: Prediction of disordered protein regions from amino acid sequence. *Nucleic Acids Research*, 35(Web Server issue), W460–W464. <https://doi.org/10.1093/nar/gkm363>
- Jin, Y., & Dunbrack, R. L. Jr. (2005). Assessment of disorder predictions in CASP6. *Proteins*, 61 Suppl 7(S7), 167–175. <https://doi.org/10.1002/prot.20734>
- Jones, D. T., & Swindells, M. B. (2002). Getting the most from PSI-BLAST. *Trends in Biochemical Sciences*, 27(3), 161–164. [https://doi.org/10.1016/s0968-0004\(01\)02039-4](https://doi.org/10.1016/s0968-0004(01)02039-4)
- Liu, Y., Wang, X., & Liu, B. (2019). A comprehensive review and comparison of existing computational methods for intrinsically disordered protein and region prediction. *Briefings in Bioinformatics*, 20(1), 330–346. <https://doi.org/10.1093/bib/bbx126>
- Melamud, E., & Moul, J. (2003). Evaluation of disorder predictions in CASP5. *Proteins*, 53 Suppl 6(S6), 561–565. <https://doi.org/10.1002/prot.10533>
- Meng, F., Uversky, V. N., & Kurgan, L. (2017). Comprehensive review of methods for prediction of intrinsic disorder and its molecular functions. *Cellular and Molecular Life Sciences: CMLS*, 74(17), 3069–3090. <https://doi.org/10.1007/s00018-017-2555-4>
- Metallo, S. J. (2010). Intrinsically disordered proteins are potential drug targets. *Current Opinion in Chemical Biology*, 14(4), 481–488. <https://doi.org/10.1016/j.cbpa.2010.06.169>
- Midic, U., Oldfield, C. J., Dunker, A. K., Obradovic, Z., & Uversky, V. N. (2009). Protein disorder in the human diseasome: Unfoldomics of human genetic diseases. *BMC Genomics*, 10 Suppl 1(Suppl 1), S12. <https://doi.org/10.1186/1471-2164-10-S1-S12>
- Mohan, A., Uversky, V. N., & Radivojac, P. (2009). Influence of sequence changes and environment on intrinsically disordered proteins. *PLoS Computational Biology*, 5(9), e1000497. <https://doi.org/10.1371/journal.pcbi.1000497>
- Monastyrskyy, B., Kryshchak, A., Moul, J., Tramontano, A., & Fidelis, K. (2014). Assessment of protein disorder region predictions in CASP10. *Proteins*, 82 Suppl 2(0 2), 127–137. <https://doi.org/10.1002/prot.24391>
- Necci, M., Piovesan, D., & Tosatto, S. C. E., CAID Predictors, DisProt Curators, Tosatto SCE. (2021). Critical assessment of protein intrinsic disorder prediction. *Nature Methods*, 18(5), 472–481. <https://doi.org/10.1038/s41592-021-01117-3>
- Noivirt-Brik, O., Prilusky, J., & Sussman, J. L. (2009). Assessment of disorder predictions in CASP8. *Proteins*, 77 Suppl 9(S9), 210–216. <https://doi.org/10.1002/prot.22586>
- Pal, M. K., Lahiri, T., & Kumar, R. (2020). ProtPCV: A fixed dimensional numerical representation of protein sequence to significantly reduce sequence search time. *Interdisciplinary Sciences, Computational Life Sciences*, 12(3), 276–287. <https://doi.org/10.1007/s12539-020-00380-w>
- Raychaudhuri, S., Dey, S., Bhattacharyya, N. P., & Mukhopadhyay, D. (2009). The role of intrinsically unstructured proteins in neurodegenerative diseases. *PloS One*, 4(5), e5566. <https://doi.org/10.1371/journal.pone.0005566>
- Receveur-Bréchet, V., Bourhis, J. M., Uversky, V. N., Canard, B., & Longhi, S. (2006). Assessing protein disorder and induced folding. *Proteins*, 62(1), 24–45. <https://doi.org/10.1002/prot.20750>
- Tripathi, A., Mondal, R., Lahiri, T., Chaurasiya, D., & Pal, M. K. (2023). TemPred: A novel protein template search engine to improve protein structure prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 20(3), 2112–2121. <https://doi.org/10.1109/tcbb.2022.3233846>
- Uversky, V. N. (2019). Intrinsically disordered proteins and their “mysterious” (meta)physics. *Frontiers in Physics*, 7. <https://doi.org/10.3389/fphys.2019.00010>
- Uversky, V. N., Oldfield, C. J., & Dunker, A. K. (2008). Intrinsically disordered proteins in human diseases: Introducing the D2 concept. *Annual Review of Biophysics*, 37(1), 215–246. <https://doi.org/10.1146/annurev.biophys.37.032807.125924>
- Uversky, V. N., Oldfield, C. J., Midic, U., Xie, H., Xue, B., Vucetic, S., Iakoucheva, L. M., Obradovic, Z., & Dunker, A. K. (2009). Unfoldomics of human diseases: Linking protein intrinsic disorder with diseases. *BMC Genomics*, 10 Suppl 1(Suppl 1), S7. <https://doi.org/10.1186/1471-2164-10-S1-S7>
- van der Lee, R., Buljan, M., Lang, B., Weatheritt, R. J., Daughdrill, G. W., Dunker, A. K., Fuxreiter, M., Gough, J., Gsponer, J., Jones, D. T., Kim, P. M., Kriwacki, R. W., Oldfield, C. J., Pappu, R. V., Tompa, P., Uversky, V. N., Wright, P. E., & Babu, M. M. (2014). Classification of intrinsically disordered regions and proteins. *Chemical Reviews*, 114(13), 6589–6631. <https://doi.org/10.1021/cr400525m>
- Walsh, I., Giollo, M., Di Domenico, T., Ferrari, C., Zimmermann, O., & Tosatto, S. C. E. (2015). Comprehensive large-scale assessment of intrinsic protein disorder. *Bioinformatics (Oxford, England)*, 31(2), 201–208. <https://doi.org/10.1093/bioinformatics/btu625>
- Yang, Z. R., Thomson, R., McNeil, P., & Esnouf, R. M. (2005). RONN: The basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics (Oxford, England)*, 21(16), 3369–3376. <https://doi.org/10.1093/bioinformatics/bti534>