Computational Prediction of Intrinsic Disorder in Protein Sequences with the disCoP Meta-predictor

Christopher J. Oldfield[1], Xiao Fan[2], Chen Wang[3], A. Keith Dunker[4] and Lukasz Kurgan[1]*

[1]Department of Computer Science, Virginia Commonwealth University, Richmond, VA 23284, USA

[2]Department of Pediatrics, Columbia University, New York, NY, 10032, USA

[3]Department of Medicine, Columbia University, New York, NY, 10032, USA

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

*Corresponding author: lkurgan@vcu.edu

Abstract

Intrinsically disordered proteins are either entirely disordered or contain disordered regions in their native state. These proteins and regions function without the prerequisite of a stable structure and were found to be abundant across all kingdoms of life. Experimental annotation of disorder lags behind the rapidly growing number of sequenced proteins, motivating the development of computational methods that predict disorder in protein sequences. DisCoP is a user-friendly webserver that provides accurate sequence-based prediction of protein disorder. It relies on meta-architecture in which the outputs generated by multiple disorder predictors are combined together to improve predictive performance. The architecture of disCoP is presented and its accuracy relative to several other disorder predictors is briefly discussed. We describe usage of the web interface and explain how to access and read results generated by this computational tool. We also provide an example of prediction results and interpretation. The disCoP's webserver is publicly available at http://biomine.cs.vcu.edu/servers/disCoP/.

Keywords

Intrinsically disordered proteins; IDP; bioinformatics; webserver; meta-architecture.

1. Introduction

Intrinsically disordered proteins (IDPs) form broad structural ensembles and lack stable folded structure in isolation under physiological conditions [16, 18, 29, 46, 86, 95]. These proteins have also been called partially folded, natively denatured, natively unfolded, natively disordered, intrinsically unstructured, intrinsically denatured, and intrinsically unfolded [16]. IDPs have one or more intrinsically disordered regions (IDRs) and in some cases they are fully disordered. Recent computational studies estimate that eukaryotic organisms have between 3% and 17% of fully disordered proteins, and that between 30% and 50% of proteins in their proteomes have at least one long IDR (30 or more consecutive amino acid residues long) [19, 68, 73, 78, 94, 101, 108]. IDPs also occupy a large part of proteomes in bacteria, archaea and viruses [5, 19, 24, 50, 78, 92, 94, 99, 101,

102, 104]. They are instrumental for numerous cellular functions including signaling [17, 22, 28, 88], regulation of transcription [26, 48], translation [75], chromatin condensing [20, 53, 74, 83], and molecular interactions with proteins and nucleic acids [6, 9, 10, 21, 27, 31, 76, 85, 97], to name just a few. Intrinsic disorder was shown to be enriched in alternatively spliced regions [3, 40, 82, 111] and in post-translational modification sites [43, 98, 111]. Moreover, IDPs are being explored as drug targets [8, 35], which is motivated by their association with a number of human diseases [57, 87].

Sequences of IDRs are substantially different from the sequences of structured regions and proteins. For example, IDRs are enriched in polar amino acids, depleted in large hydrophobic and aromatic amino acid, and have relatively low sequence complexity [4, 45, 80, 81]. These differences underlie the development of accurate computational methods for the prediction of disorder in protein chains. Over 70 computational disorder predictors were developed over the last few decades [2, 11, 14, 15, 25, 32, 44, 47, 54, 55, 65, 66, 79]. Many of the recently published methods rely on meta-architectures that combine outputs produced by several disorder predictors to (re)predict disorder. The meta-predictors include (in chronological order) VSL2 [70], metaPrDOS [37], PreDisorder [12], NN-CDF [103], MD [84], PONDR-FIT [100], MFDp [61], CSpritz [90], MetaDisorder [41], ESpritz [89], MFDp2 [60, 63], DisMeta [36], disCoP [23], DISOPRED3 [38], and MobiDB-lite [67]. This type of predictive architecture is motivated by studies that empirically demonstrate that outputs from the meta-predictors are more accurate when compared to the results produced by their input single predictors [23, 72]. However, the improved accuracy comes at a cost of a longer runtime and inconvenience. The long runtime stems from the fact that multiple disorder predictions have to be computed and combined together. The inconvenience is due to the fact that outputs of several disorder predictors must be collected by the user. The latter drawback is alleviated by some meta-predictors that incorporate computation of the input disorder predictors into their publicly available implementations.

A recently published example of a convenient meta-predictor is disCoP (**dis**order **Co**nsensus-based **P**redictor) [23]. The disCoP method is available as a user-friendly webserver that automates the entire prediction process. Users only need to enter the sequence of their proteins and click the “Run” button to obtain disorder prediction. Moreover, benchmarking tests show that DisCoP provides accurate predictions, with area under the receiver operating characteristic (ROC) curve (AUC) = 0.85 and Matthews correlation coefficient (MCC) = 0.50. DisCoP was compared empirically to 20 other disorder predictors including several meta-predictors such as ESpritz (AUC = 0.83 and MCC = 0.48), CSpritz (AUC = 0.83 and MCC = 0.45), MD (AUC = 0.82 and MCC = 0.45), MFDp (AUC = 0.82 and MCC = 0.45) and PONDR-FIT (AUC = 0.78 and MCC = 0.41). These tests concluded that predictive performance of disCoP is statistically significantly better (p-value < 0.01) [23]. To sum up, the two main advantages of disCoP are the availability of the convenient webserver and good predictive performance.

This chapter describes the underlying meta-architecture of disCoP, explains its web interface and provides detailed instructions on how to generate predictions with this computational tool. We also explain how to read and interpret the results generated by this meta-predictor using a case study that concerns prediction of intrinsic disorder for the chromatin accessibility complex 16kD protein.

2. Materials

1. Sequences of proteins to be predicted. The sequences must be formatted using the FASTA format (*see* **Note 1**). Up to 5 protein sequences can be submitted at one time as either a file upload or using a text entry field (*see* **Note 2**).

2. disCoP: The webserver that is freely available at http://biomine.cs.vcu.edu/servers/disCoP/ is designed to be simple to use. All computations are performed on the server side and thus the only requirements for submitting predictions are: an internet connection and a modern web browser (Firefox, Internet Explorer, or Chrome). The webserver visualizes the results directly in the web browser window and also delivers these results to the user-provided email address.

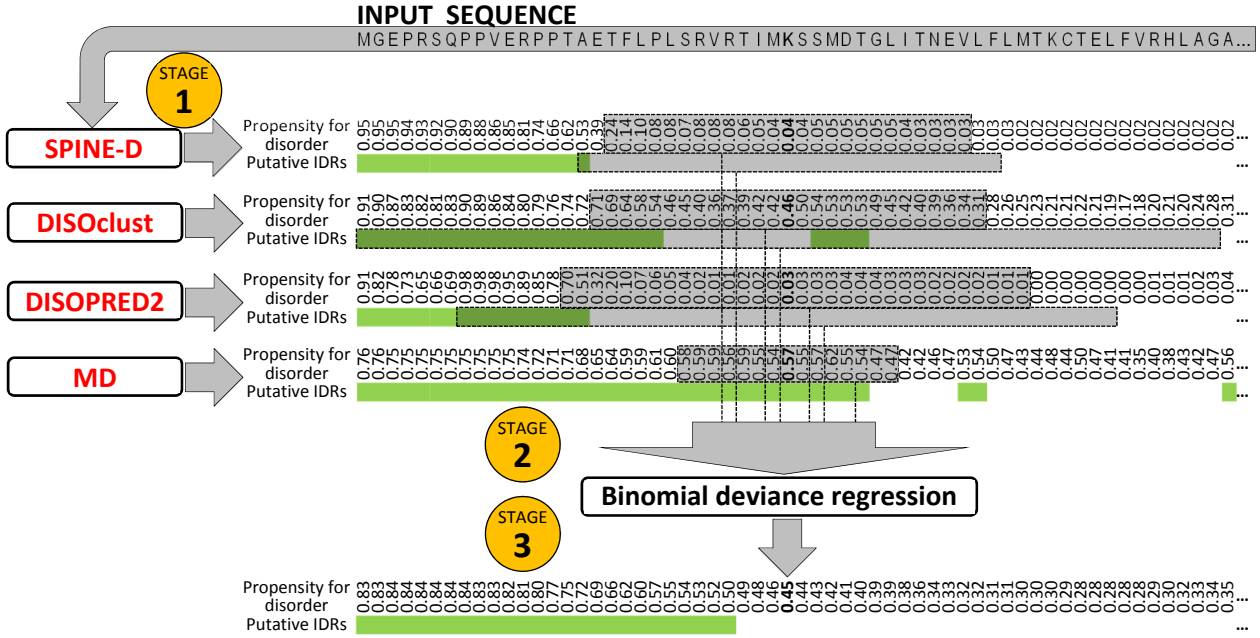

Fig. 1 Prediction process implemented in the disCoP predictor. The outputs of the four disorder predictors (SPINE-D, DISOclust, DISOPRED2 and MD) generated in stage 1 include the propensity scores and the corresponding putative IDRs, which are show using the green horizontal bars. The dashed boxes with gray shading denote the sliding windows that are used to compute the seven features in stage 2. In stage 3, the binomial deviation regression model predicts the putative propensities for disorder from the seven features. The putative IDRs generated by disCOP are shown at the bottom of the figure and they correspond to the residues for which the putative propensities for disorder ≥ 0.5. The example shows results produced for the chromatin accessibility complex 16kD protein (UniProt id: Q9V452)

The meta-architecture of the disCoP's webserver is shown in Fig. 1. The input protein sequence goes through a three-stage process to generate putative IDRs. In stage 1, the sequence is processed by four disorder predictors: SPINE-D [110], DISOclust [49], DISOPRED2 [93] and MD [84]. This collection of four predictors was selected from among 20 disorder predictors using an empirical procedure that aims to maximize predictive performance [23]. Each of the four methods outputs numeric propensity for disorder and binary disorder annotations (disordered vs. ordered) for each residue in the input protein chain. In stage 2, these predictions are processed to produce features that numerically quantify information which is relevant for the disorder prediction. The features are calculated using sliding windows that aggregate and summarize putative disorder information among neighboring (in the sequence) amino acid residues. This reduces the risk of making spurious predictions. The windows are represented by dashed boxes in Fig. 1. A balanced and complementary set of seven features is collected by considering both types of outputs (propensities and binary) generated by each of the four disorder predictors. Stage 3 uses these features as input to a trained regression model to produce disCoP's predictions in the form of numeric propensities for disorder. These propensities range between 0 and 1, where higher propensity scores are indicative of a higher likelihood of intrinsic disorder. The disCoP's webserver further processes these propensities to generate binary predictions,

which correspond to the putative IDRs. Residues with propensities > 0.5 are predicted to be disordered while the remaining residues are predicted as ordered/structured (*see* **Note 3**).

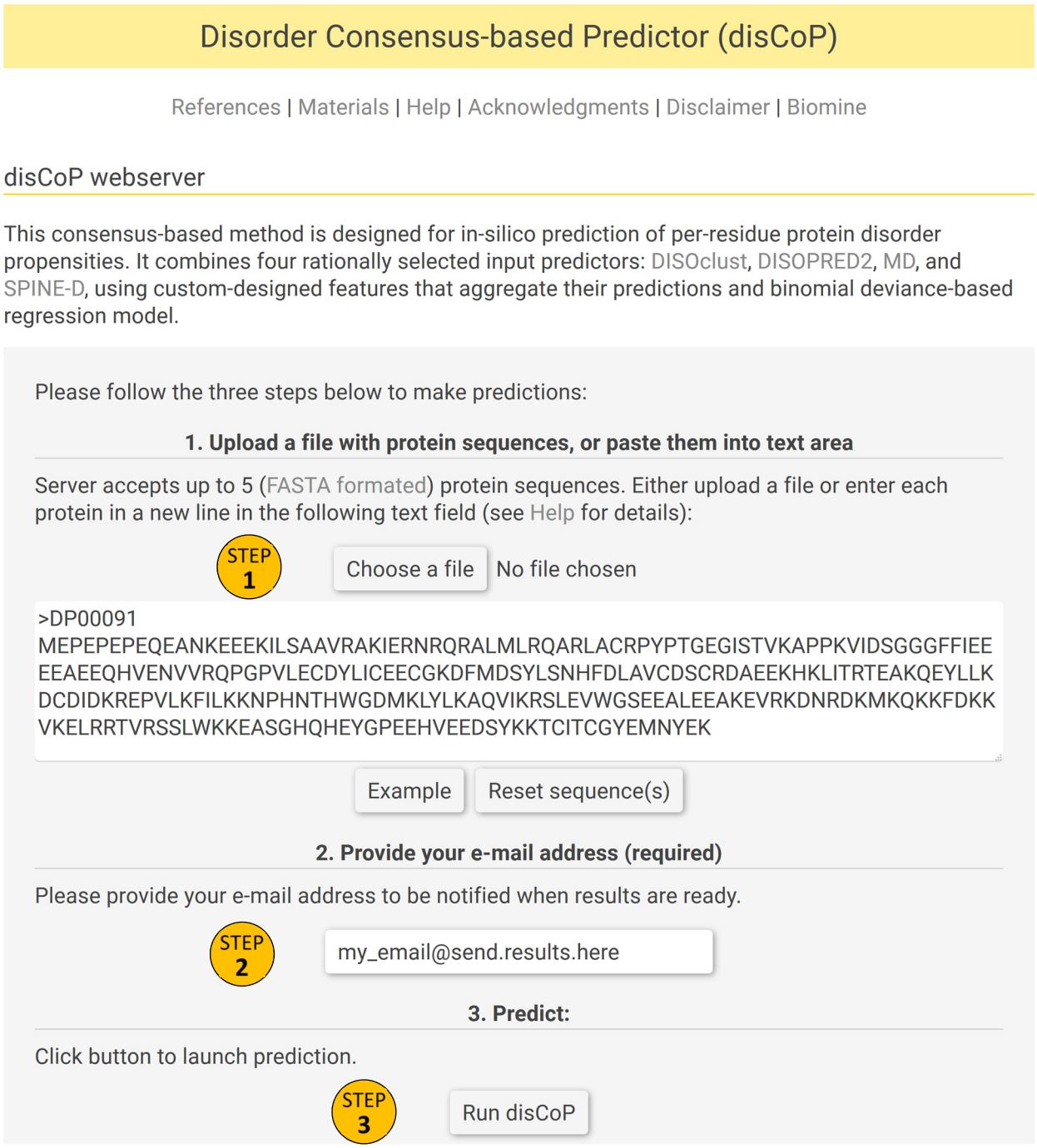

Fig. 2. The disCoP prediction submission webpage. Orange/yellow circles indicate the three steps to submit sequences for predictions, discussed in the text.

3. Methods

Submission of predictions is made at the main disCoP's webpage at http://biomine.cs.vcu.edu/servers/disCoP/. Notification of completed predictions are given by email,

and thus an email address is required for each submission. These notifications provide a link to prediction results, which can be viewed in a browser window and/or downloaded as a parsable text file. The predictions can be accessed at a later time and they are kept on the webserver for at least three months.

3.1. Running disCoP

Three easy steps are required to submit sequences for prediction (Fig. 2, labels 1, 2, and 3):

Step 1. Enter FASTA formatted sequences (*see* **Note 1**) in one of two ways:

- Upload a file of FASTA formatted sequences.
- Input the FASTA formatted sequences into the white text entry field. This can be done using the copy and paste function. An example of properly formatted sequence can be obtained by clicking the "Example" button located below the text entry field.

Clicking the "Reset sequence(s)" button clears both submission options. There are limits to both the number of sequences and maximum length of sequences that can be submitted for prediction (*see* **Note 2**).

Step 2. Provide an email address (*see* **Note 4**). This email is only used to send notification of completed predictions.

Step 3. Click "Run disCoP" to start the prediction.

Clicking "Run disCoP" takes the user to a status page that reports on the current state of the submitted prediction. Submissions to several different bioinformatics webservers located at the http://biomine.cs.vcu.edu site (*see* **Note 5**) are entered into the same queue system (*see* **Note 6**). The status page reports the current position in the queue and shows when prediction for this submission begins. The runtime needed to complete prediction for an average length protein sequence (about 250 amino acids) is approximately 10 minutes. The prediction can take over 40 minutes when submitting 5 longer protein sequences. After the prediction is completed, the status page automatically redirects the user to the prediction results page. This also triggers an email with the location of the results page that is sent to the user-provided email address. There is no need to keep the status page open while predictions are running since the notification email is always sent when the prediction is finished.

Predictions for disCoP job id: 20190404185835 are ready. 1

Upon the usage the users are requested to use the following citation(s):
Fan X, Kurgan LA, 2014. Accurate prediction of disorder in protein chains with a comprehensive and empirically designed consensus. Journal of Biomolecular Structure and Dynamics, 32(3): 448-464.

You can find the results for this job at:
http://biomine.cs.vcu.edu/webresults/disCoP/20190404185835/results.html 2

The CSV file can be found here: http://biomine.cs.vcu.edu/webresults/disCoP/20190404185835/results.csv

The webserver can be found here: http://biomine.cs.vcu.edu/servers/disCoP/ 3

Thank you for using our webserver,
Biomine group

Fig. 3. The disCoP notification email. The email provides unique job identifier and links to the results indicated with orange/yellow circles, discussed in the text.

3.2. Results generated by disCoP

The results page can be reached by leaving the status page open for the duration of the prediction, or by following the link provided in the email.

The email (Fig. 3) provides a job identifier together with the location of the prediction results page and the text file with the results. Each submission is assigned a unique 14-digits long identifier (Fig. 3, label 1) that is shown at the top of the notification email (*see* **Note 7**). In the case of issues with the completion and/or contents of the prediction, the identifier can be used to trace the corresponding submission. The email message includes a direct link to the webpage with the results of the prediction (Fig. 3, label 2) and to the text file with the results (Fig. 3, label 3).

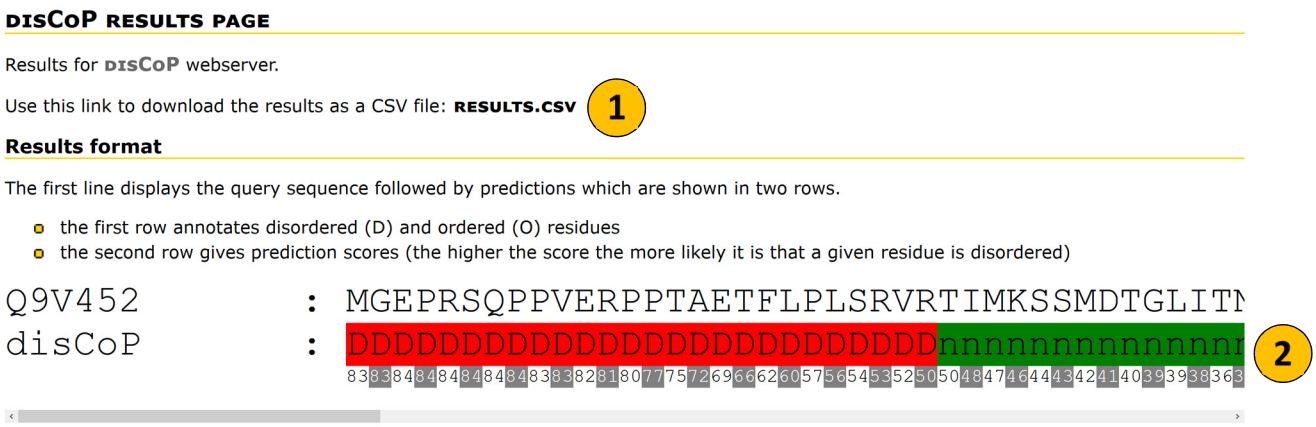

Fig. 4. The disCoP prediction results webpage. The red D and green n denote the putative disordered residues and putative non-disordered (structured) residues, respectively. The corresponding putative propensity scores are provided directly underneath. Orange/yellow circles indicate important features of this page, discussed in the text.

This results page (Fig. 4) includes a link to a text file *results.csv* with prediction results (Fig. 4, label 1) and a visualization of the predictions (Fig. 4, label 2) (*see* **Note 8**). The text file contains protein identifiers, sequences, binary predictions and propensity scores for each of the submitted protein sequences. These data are in the comma-separable CSV format to ease parsing. An example of the CSV format results file is shown in Fig. 5. Each sequence is represented by three lines:

Line 1. The protein name taken from the FASTA header provided by the user followed by the protein sequence. The individual amino acids are comma-separated to ease aligning them to the corresponding predictions listed in the two subsequent lines.

Line 2. Binary predictions of disorder, where D denotes disordered residues and O denoted ordered residues.

Line 3. Propensity for disorder, ranging between 1 for high propensity to 0 for low propensity. Residues with propensity > 0.5 are annotated as D in the second line.

The visualization (Fig. 4, label 2) shows the binary annotations of the putative IDRs (using red highlights) and putative ordered regions (green highlights) for each residue in the input protein chain. Each binary annotation is associated with the propensity score, which is provided directly underneath. The scores are in the range between 0 and 99, where residues with scores > 50 are predicted in binary as disordered.

```
Q9V452,M,G,E,P,R,S,Q,P,P,V,E,R,P,P,T,A,E,T,F,L,P,L,S,R,V,R,T,I,M,K,S,S,M,D,T,G,L,I,T,N,E,V,L,
F,L,M,T,K,C,T,E,L,F,V,R,H,L,A,G,A,A,Y,T,E,E,F,G,Q,R,P,G,E,A,L,K,Y,E,H,L,S,Q,V,V,N,K,N,K,N,L,E
,F,L,L,Q,I,V,P,Q,K,I,R,V,H,Q,F,Q,E,M,L,R,L,N,R,S,A,G,S,D,D,D,D,D,D,D,D,D,D,E,E,E,S,E,S,E,S,
E,S,D,E
disCoP,D,D,D,D,D,D,D,D,D,D,D,D,D,D,D,D,D,D,D,D,D,D,D,D,D,D,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,
O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,
,O,O,O,O,O,O,O,O,O,O,O,O,D,D,D,D,D,D,D,D,D,D,D,D,D,D,D,D,D,D,D,D,D,D,D,D,D,D,D,D,D,D,D,D,D,
D,D,D,D
disCoP - prob,0.832,0.832,0.835,0.838,0.840,0.840,0.840,0.837,0.834,0.829,0.824,0.810,0.795,0
.770,0.749,0.721,0.689,0.655,0.623,0.600,0.577,0.558,0.542,0.529,0.519,0.507,0.496,0.479,0.46
8,0.458,0.443,0.431,0.416,0.408,0.399,0.393,0.388,0.376,0.359,0.342,0.328,0.317,0.315,0.314,0
.310,0.304,0.300,0.295,0.289,0.284,0.278,0.275,0.277,0.284,0.293,0.302,0.317,0.329,0.337,0.34
9,0.361,0.377,0.396,0.413,0.422,0.430,0.436,0.438,0.441,0.442,0.442,0.440,0.437,0.435,0.431,0
.426,0.419,0.409,0.399,0.393,0.385,0.378,0.363,0.353,0.344,0.334,0.323,0.323,0.322,0.324,0.33
1,0.336,0.342,0.349,0.360,0.389,0.415,0.451,0.463,0.478,0.496,0.516,0.535,0.556,0.572,0.587,0
.605,0.626,0.650,0.678,0.711,0.747,0.761,0.768,0.771,0.770,0.769,0.767,0.762,0.758,0.753,0.74
9,0.744,0.744,0.745,0.744,0.744,0.744,0.744,0.744,0.740,0.733,0.726,0.717,0.711,0.706,0.706,0
.706,0.706,0.706
```

Fig. 5. Example of the CSV format results file for the disCoP prediction. The example shows results produced for the chromatin accessibility complex 16kD protein (UniProt id: Q9V452).

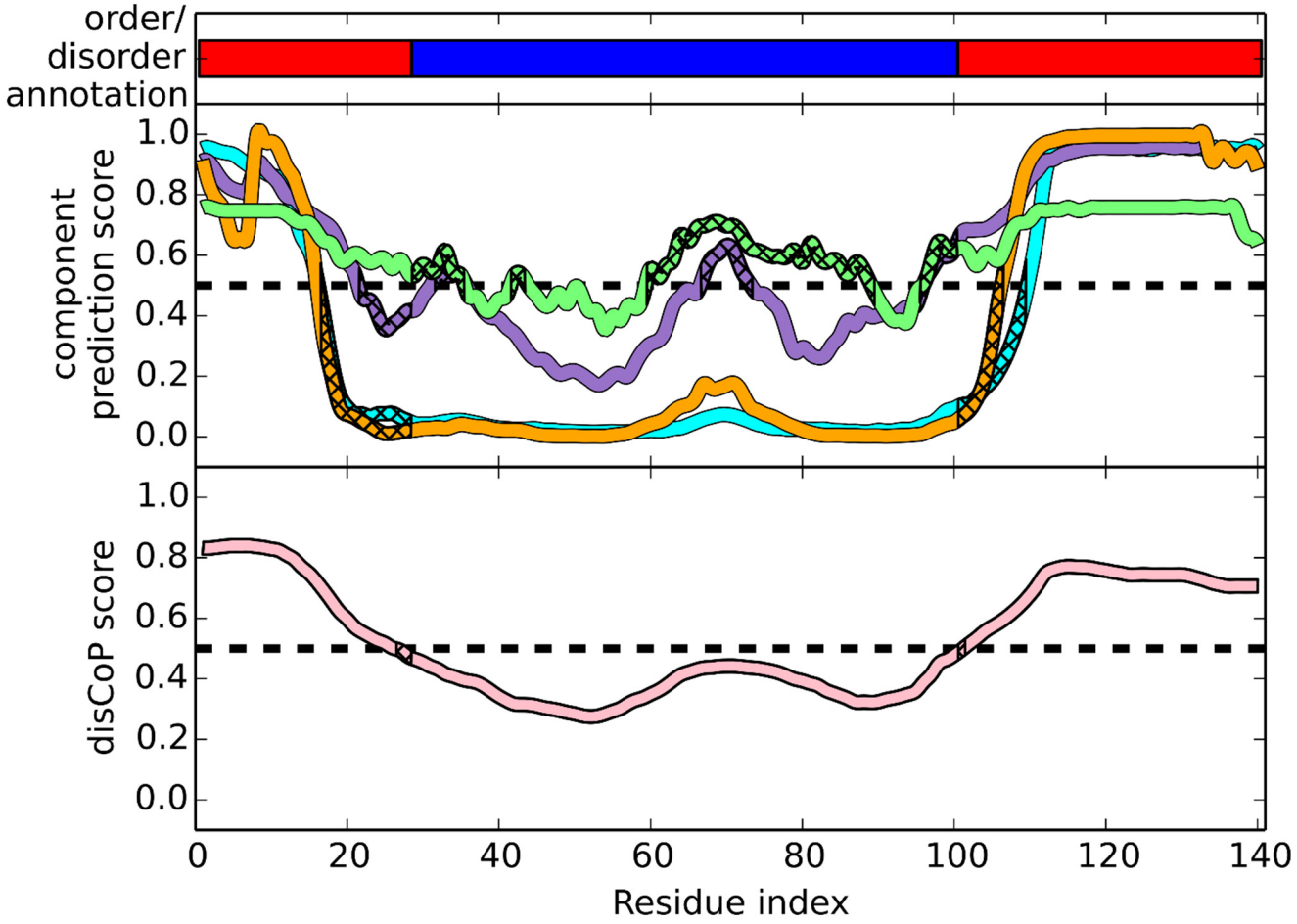

Fig. 6. Known intrinsically disordered and structured regions in CHRAC16 compared to disorder predictions. (Top panel) Structurally characterized regions are shown: two intrinsically disordered regions (red) and one structured region (blue). (Middle and bottom panels) Intrinsic disorder prediction scores given by SPINE-D (cyan), DISOPRED (orange), DISOclust (green), MD (purple), and disCoP (pink, shown alone in the bottom panel) are shown, where values above 0.5 are predictions of disorder and below 0.5 are prediction of structure. Hatch portions of the score lines indicate incorrect predictions.

4. Case study

The protein CHRAC16 is a component of the chromatin accessibility complex (CHRAC), formed by interaction of CHRAC16 and CHRAC14 with the ATP-utilizing chromatin assembly and remodeling factor (ACF) complex. The crystal structure of the CHRAC14/16 dimer has been determined [30], which revealed two disordered regions located at either terminus (Fig. 6, top panel). The N- and C-terminal IDRs play a role in the ACF binding and modulating DNA binding affinity, respectively [30].

Disorder predictions for CHRAC16 demonstrates the improvement of disCoP predictions relative to its component predictions from SPINE-D, DISOPRED, DISOclust, and MD. This is shown in Fig. 6 by comparing the amount of incorrect predictions (hatch portions of the score lines) between disCoP and the other four methods. For comparison, prediction scores from disCoP server's CSV output file were plotted along with prediction scores of the component predictors (Fig. 6, middle and bottom panels). The four component predictors of disCoP generally perform well for Drosophila CHRAC16 (Fig. 6, middle panel); SPINE-D, DISOPRED, DISOclust, and MD predict 85%, 87%, 84% and 69% of residues correctly, respectively. Both SPINE-D and DISOPRED predict disordered and ordered regions correctly, but predict the two disordered regions to be shorter than found experimentally. DISOclust and MD both predict too much disorder, with MD predicting much of the structured region to be disordered. In contrast, the disCoP prediction is highly accurate (Fig. 6, bottom panel), predicting 98% of residues correctly. Similar to SPINE-D and DISOPRED, disCoP slightly under predicts disorder at the N-terminus and C-terminus, but only by two residues and one residues, respectively.

Notes

1. The FASTA format for the protein sequences is explained at https://en.wikipedia.org/wiki/FASTA_format. Briefly, each protein is represented by multiple lines where the first line that begins with ">" followed by the name and description of the protein, and the subsequent lines that provide the sequence using the 1-letter amino acid encoding and with 20 amino acids per line. Example follows

   ```
   >Q9V452
   MGEPRSQPPVERPPTAETFLPLSRVRTIMKSSMDTGLITNEVLFLMTKCTELFVRHLAGA
   AYTEEFGQRPGEALKYEHLSQVVNKNKNLEFLLQIVPQKIRVHQFQEMLRLNRSAGSDDD
   DDDDDDDDEEESESESESDE
   ```

   The disCoP server will also accept the second line that gives the entire protein sequence, i.e., the user has the option of providing the sequence in one line or breaking it up into multiple lines.
2. Up to five FASTA-formatted sequences can be submitted at one time. Moreover, the programs used to implement the disCoP predictor limit the length of submitted protein sequences to the range between 26 residues and 1000 residues. These limits apply to both the text entry field and when uploading the file. Submissions exceeding either of these limits receive an error notification from the server ("You entered 10 proteins. Up to 5 proteins allowed!" or "Input sequence is 1024 amino acids long. The minimal allowed length is 26 amino acids and the maximal length is 1000. Please re-submit your sequence.") and prediction is disallowed. Requests with more than 5 proteins have to be broken into multiple submissions each with 5 or fewer sequences; (also, *see* **Note 6**). The users must combine the results from different submissions manually.
3. Analysis of the predictions generated by disCoP benefits from examining the propensity scores in addition to the binary predictions. High values of the propensity scores which are below the 0.5

threshold (and which consequently do not result in the binary prediction of IDRs) may suggest presence of disorder if combined with other data. Benchmarks show that the threshold = 0.5 corresponds to the predictions with sensitivity of about 65% and low (15%) false positive rate, resulting in a rather conservative set of disorder predictions. This means that residues that were not predicted as disordered based on the binary outputs and which have high propensity scores have elevated likelihood for disorder, but at higher levels of false positives.

4. Rather than requiring an active browser connection for the duration of the entire prediction, notification of completed predictions are provided via the email address provided by the user.
5. The http://biomine.cs.vcu.edu site includes several other predictors, such as (in alphabetical order) CONNECTOR [91], CRYSTALP2 [42], Cypred [39], DFLpred [51], DisCon [64], DisoRDPbind [71, 77], DMRpred [52], DRNApred [106], fDETECT [56, 58], fMoRFpred [105], funDNApred [1], hybridNAP [109], ILbind [33], MFDp [62], MFDp2 [60, 63], MoRFpred [13, 69], NsitePred [7], PPCpred [59], QUARTER [34, 96], RAPID [108], SSCon [107] and SLIDER [73].
6. The http://biomine.cs.vcu.edu site utilizes the first-come-first-serve queue. However, the number of simultaneous submissions across all webservers (*see* **Note 5**) that are received from the same IP address is limited to three. Users who submit too frequently receive a message to resubmit after one of their pending submissions is completed. This limit aims to equalize access to this resource across users by not allowing any one user to submit an excessive number of jobs that would severely delay/block access for the other users.
7. Both links to the results are based on the unique job identifier and they are not posted online. This means that the other users of this webserver are unable to access the results, preserving privacy of the submission.
8. Users should save the email and the links to the results. They can be accessed only via the links that are provided in the notification email and on the results webpage.

Acknowledgements

This research was supported in part by the Robert J. Mattauch Endowment funds and the National Science Foundation grant 1617369 to Lukasz Kurgan.

References

1. Amirkhani A, Kolahdoozi M, Wang C et al. (2018) Prediction of DNA-binding residues in local segments of protein sequences with Fuzzy Cognitive Maps. IEEE/ACM Trans Comput Biol Bioinform
2. Atkins J, Boateng S, Sorensen T et al. (2015) Disorder Prediction Methods, Their Applicability to Different Protein Targets and Their Usefulness for Guiding Experimental Studies. International Journal of Molecular Sciences 16:19040
3. Buljan M, Chalancon G, Dunker AK et al. (2013) Alternative splicing of intrinsically disordered regions and rewiring of protein interactions. Current opinion in structural biology 23:443-450
4. Campen A, Williams RM, Brown CJ et al. (2008) TOP-IDP-scale: a new amino acid scale measuring propensity for intrinsic disorder. Protein Pept Lett 15:956-963
5. Charon J, Theil S, Nicaise V et al. (2016) Protein intrinsic disorder within the Potyvirus genus: from proteome-wide analysis to functional annotation. Molecular Biosystems 12:634-652
6. Chen JW, Romero P, Uversky VN et al. (2006) Conservation of intrinsic disorder in protein domains and families: II. Functions of conserved disorder. Journal of Proteome Research 5:888-898
7. Chen K, Mizianty MJ, Kurgan L (2012) Prediction and analysis of nucleotide-binding residues using sequence and sequence-derived structural descriptors. Bioinformatics 28:331-341

8. Cheng Y, Legall T, Oldfield CJ et al. (2006) Rational drug design via intrinsically disordered protein. Trends Biotechnol 24:435-442
9. Chowdhury S, Zhang J, Kurgan L (2018) In Silico Prediction and Validation of Novel RNA Binding Proteins and Residues in the Human Proteome. Proteomics:e1800064
10. Cumberworth A, Lamour G, Babu MM et al. (2013) Promiscuity as a functional trait: intrinsically disordered regions as central players of interactomes. Biochemical Journal 454:361-369
11. Deng X, Eickholt J, Cheng J (2012) A comprehensive overview of computational protein disorder prediction methods. Mol Biosyst 8:114-121
12. Deng X, Eickholt J, Cheng J (2009) PreDisorder: ab initio sequence-based prediction of protein disordered regions. BMC Bioinformatics 10:436
13. Disfani FM, Hsu WL, Mizianty MJ et al. (2012) MoRFpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins. Bioinformatics 28:i75-83
14. Dosztányi Z, Mészáros B, Simon I (2010) Bioinformatical approaches to characterize intrinsically disordered/unstructured proteins. Briefings in bioinformatics 11:225-243
15. Dosztányi Z, Tompa P (2008) Prediction of Protein Disorder. In: Kobe B, Guss M, Huber T (eds) Structural Proteomics. Humana Press, p 103-115
16. Dunker AK, Babu MM, Barbar E et al. (2013) What's in a name? Why these proteins are intrinsically disordered. Intrinsically Disordered Proteins 1:e24157
17. Dunker AK, Cortese MS, Romero P et al. (2005) Flexible nets. The roles of intrinsic disorder in protein interaction networks. FEBS J 272:5129-5148
18. Dunker AK, Obradovic Z (2001) The protein trinity--linking function and disorder. Nature biotechnology 19:805-806
19. Dunker AK, Obradovic Z, Romero P et al. (2000) Intrinsic protein disorder in complete genomes. Genome Inform Ser Workshop Genome Inform 11:161-171
20. Dyson HJ (2012) Roles of intrinsic disorder in protein-nucleic acid interactions. Mol Biosyst 8:97-104
21. Dyson HJ (2012) Roles of intrinsic disorder in protein-nucleic acid interactions. Molecular Biosystems 8:97-104
22. Dyson HJ, Wright PE (2005) Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol 6:197-208
23. Fan X, Kurgan L (2014) Accurate prediction of disorder in protein chains with a comprehensive and empirically designed consensus. J Biomol Struct Dyn 32:448-464
24. Fan X, Xue B, Dolan PT et al. (2014) The intrinsic disorder status of the human hepatitis C virus proteome. Mol Biosyst 10:1345-1363
25. Ferron F, Longhi S, Canard B et al. (2006) A practical overview of protein disorder prediction methods. Proteins: Structure, Function, and Bioinformatics 65:1-14
26. Fuxreiter M, Tompa P, Simon I et al. (2008) Malleable machines take shape in eukaryotic transcriptional regulation. Nat Chem Biol 4:728-737
27. Fuxreiter M, Toth-Petroczy A, Kraut DA et al. (2014) Disordered Proteinaceous Machines. Chem Rev 114:6806-6843
28. Galea CA, Wang Y, Sivakolundu SG et al. (2008) Regulation of cell division by intrinsically unstructured proteins: intrinsic flexibility, modularity, and signaling conduits. Biochemistry 47:7598-7609
29. Habchi J, Tompa P, Longhi S et al. (2014) Introducing protein intrinsic disorder. Chem Rev 114:6561-6588
30. Hartlepp KF, Fernandez-Tornero C, Eberharter A et al. (2005) The histone fold subunits of Drosophila CHRAC facilitate nucleosome sliding through dynamic DNA interactions. Molecular and cellular biology 25:9886-9896
31. Haynes C, Oldfield CJ, Ji F et al. (2006) Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. PLoS Comput Biol 2:890-901
32. He B, Wang K, Liu Y et al. (2009) Predicting intrinsic disorder in proteins: an overview. Cell Res 19:929-949

33. Hu G, Gao J, Wang K et al. (2012) Finding protein targets for small biologically relevant ligands across fold space using inverse ligand binding predictions. Structure 20:1815-1822
34. Hu G, Wu Z, Oldfield C et al. (2018) Quality Assessment for the Putative Intrinsic Disorder in Proteins. Bioinformatics
35. Hu G, Wu Z, Wang K et al. (2016) Untapped Potential of Disordered Proteins in Current Druggable Human Proteome. Curr Drug Targets 17:1198-1205
36. Huang YJ, Acton TB, Montelione GT (2014) DisMeta: a meta server for construct design and optimization. Methods Mol Biol 1091:3-16
37. Ishida T, Kinoshita K (2008) Prediction of disordered regions in proteins based on the meta approach. Bioinformatics 24:1344-1348
38. Jones DT, Cozzetto D (2015) DISOPRED3: precise disordered region predictions with annotated protein-binding activity. Bioinformatics 31:857-863
39. Kedarisetti P, Mizianty MJ, Kaas Q et al. (2014) Prediction and characterization of cyclic proteins from sequences in three domains of life. Biochim Biophys Acta 1844:181-190
40. Korneta I, Bujnicki JM (2012) Intrinsic disorder in the human spliceosomal proteome. PLoS Comput Biol 8:e1002641
41. Kozlowski LP, Bujnicki JM (2012) MetaDisorder: a meta-server for the prediction of intrinsic disorder in proteins. BMC Bioinformatics 13:1-11
42. Kurgan L, Razib AA, Aghakhani S et al. (2009) CRYSTALP2: sequence-based protein crystallization propensity prediction. Bmc Structural Biology 9
43. Kurotani A, Tokmakov AA, Kuroda Y et al. (2014) Correlations between predicted protein disorder and post-translational modifications in plants. Bioinformatics 30:1095–1103
44. Li J, Feng Y, Wang X et al. (2015) An Overview of Predictors for Intrinsically Disordered Proteins over 2010–2014. International Journal of Molecular Sciences 16:23446
45. Li X, Romero P, Rani M et al. (1999) Predicting Protein Disorder for N-, C-, and Internal Regions. Genome Inform Ser Workshop Genome Inform 10:30-40
46. Lieutaud P, Ferron F, Uversky AV et al. (2016) How disordered is my protein and what is its disorder for? A guide through the "dark side" of the protein universe. Intrinsically Disord Proteins 4:e1259708
47. Lieutaud P, Ferron F, Uversky AV et al. (2016) How Disordered is My Protein and What is Its Disorder For? A Guide Through the “Dark Side” of the Protein Universe. Intrinsically Disordered Proteins 4:e1259708
48. Liu J, Perumal NB, Oldfield CJ et al. (2006) Intrinsic disorder in transcription factors. Biochemistry 45:6873-6888
49. Mcguffin LJ (2008) Intrinsic disorder prediction from the analysis of multiple protein fold recognition models. Bioinformatics 24:1798-1804
50. Meng F, Badierah RA, Almehdar HA et al. (2015) Unstructural biology of the Dengue virus proteins. FEBS J 282:3368-3394
51. Meng F, Kurgan L (2016) DFLpred: High-throughput prediction of disordered flexible linker regions in protein sequences. Bioinformatics 32:i341-i350
52. Meng F, Kurgan L (2018) High-throughput prediction of disordered moonlighting regions in protein sequences. Proteins
53. Meng F, Na I, Kurgan L et al. (2016) Compartmentalization and Functionality of Nuclear Disorder: Intrinsic Disorder and Protein-Protein Interactions in Intra-Nuclear Compartments. Int J Mol Sci 17
54. Meng F, Uversky V, Kurgan L (2017) Computational Prediction of Intrinsic Disorder in Proteins. Curr Protoc Protein Sci 88:2 16 11-12 16 14
55. Meng F, Uversky VN, Kurgan L (2017) Comprehensive review of methods for prediction of intrinsic disorder and its molecular functions. Cell Mol Life Sci 74:3069-3090
56. Meng F, Wang C, Kurgan L (2018) fDETECT webserver: fast predictor of propensity for protein production, purification, and crystallization. BMC Bioinformatics 18:580

57. Midic U, Oldfield CJ, Dunker AK et al. (2009) Unfoldomics of human genetic diseases: illustrative examples of ordered and intrinsically disordered members of the human diseasome. Protein and peptide letters 16:1533-1547
58. Mizianty MJ, Fan X, Yan J et al. (2014) Covering complete proteomes with X-ray structures: a current snapshot. Acta Crystallogr D Biol Crystallogr 70:2781-2793
59. Mizianty MJ, Kurgan L (2011) Sequence-based prediction of protein crystallization, purification and production propensity. Bioinformatics 27:i24-33
60. Mizianty MJ, Peng ZL, Kurgan L (2013) MFDp2: Accurate predictor of disorder in proteins by fusion of disorder probabilities, content and profiles. Intrinsically Disordered Proteins 1:e24428
61. Mizianty MJ, Stach W, Chen K et al. (2010) Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources. Bioinformatics 26:i489-i496
62. Mizianty MJ, Stach W, Chen K et al. (2010) Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources. Bioinformatics 26:i489-496
63. Mizianty MJ, Uversky V, Kurgan L (2014) Prediction of intrinsic disorder in proteins using MFDp2. Methods Mol Biol 1137:147-162
64. Mizianty MJ, Zhang T, Xue B et al. (2011) In-silico prediction of disorder content using hybrid sequence representation. BMC Bioinformatics 12:245
65. Monastyrskyy B, Kryshtafovych A, Moult J et al. (2014) Assessment of protein disorder region predictions in CASP10. Proteins 82 Suppl 2:127-137
66. Necci M, Piovesan D, Dosztanyi Z et al. (2017) A comprehensive assessment of long intrinsic protein disorder from the DisProt database. Bioinformatics
67. Necci M, Piovesan D, Dosztanyi Z et al. (2017) MobiDB-lite: fast and highly specific consensus prediction of intrinsic disorder in proteins. Bioinformatics 33:1402-1404
68. Oates ME, Romero P, Ishida T et al. (2013) D(2)P(2): database of disordered protein predictions. Nucleic Acids Res 41:D508-516
69. Oldfield CJ, Uversky VN, Kurgan L (2018) Predicting Functions of Disordered Proteins with MoRFpred. Methods Mol Biol 1851
70. Peng K, Radivojac P, Vucetic S et al. (2006) Length-dependent prediction of protein intrinsic disorder. BMC Bioinformatics 7:208
71. Peng Z, Kurgan L (2015) High-throughput prediction of RNA, DNA and protein binding regions mediated by intrinsic disorder. Nucleic Acids Res 43:e121
72. Peng Z, Kurgan L (2012) On the complementarity of the consensus-based disorder prediction. Pac Symp Biocomput:176-187
73. Peng Z, Mizianty MJ, Kurgan L (2014) Genome-scale prediction of proteins with long intrinsically disordered regions. Proteins 82:145-158
74. Peng Z, Mizianty MJ, Xue B et al. (2012) More than just tails: intrinsic disorder in histone proteins. Mol Biosyst 8:1886-1901
75. Peng Z, Oldfield CJ, Xue B et al. (2014) A creature with a hundred waggly tails: intrinsically disordered proteins in the ribosome. Cell Mol Life Sci 71:1477-1504
76. Peng Z, Sakai Y, Kurgan L et al. (2014) Intrinsic disorder in the BK channel and its interactome. PLoS One 9:e94331
77. Peng Z, Wang C, Uversky VN et al. (2017) Prediction of Disordered RNA, DNA, and Protein Binding Regions Using DisoRDPbind. Methods Mol Biol 1484:187-203
78. Peng Z, Yan J, Fan X et al. (2015) Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life. Cell Mol Life Sci 72:137-151
79. Pentony M, Ward J, Jones D (2010) Computational Resources for the Prediction and Analysis of Native Disorder in Proteins. In: Hubbard SJ, Jones AR (eds) Proteome Bioinformatics. Humana Press, p 369-393
80. Romero P, Obradovic Z, Kissinger C et al. (1997) Identifying disordered regions in proteins from amino acid sequence. In: Neural Networks,1997., International Conference on. p 90-95 vol.91

81. Romero P, Obradovic Z, Li X et al. (2001) Sequence complexity of disordered protein. Proteins: Structure, Function, and Bioinformatics 42:38-48
82. Romero PR, Zaidi S, Fang YY et al. (2006) Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms. Proc. Natl. Acad. Sci. U. S. A. 103:8390-8395
83. Sandhu KS (2009) Intrinsic disorder explains diverse nuclear roles of chromatin remodeling proteins. J Mol Recognit 22:1-8
84. Schlessinger A, Punta M, Yachdav G et al. (2009) Improved disorder prediction by combination of orthogonal approaches. PLoS One 4:e4433
85. Tompa P, Csermely P (2004) The role of structural disorder in the function of RNA and protein chaperones. Faseb Journal 18:1169-1175
86. Uversky VN, Gillespie JR, Fink AL (2000) Why are "natively unfolded" proteins unstructured under physiologic conditions? Proteins 41:415-427
87. Uversky VN, Oldfield CJ, Dunker AK (2008) Intrinsically disordered proteins in human diseases: introducing the D2 concept. Annu Rev Biophys 37:215-246
88. Uversky VN, Oldfield CJ, Dunker AK (2005) Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling. J Mol Recognit 18:343-384
89. Walsh I, Martin AJM, Di Domenico T et al. (2012) ESpritz: accurate and fast prediction of protein disorder. Bioinformatics 28:503-509
90. Walsh I, Martin AJM, Di Domenico T et al. (2011) CSpritz: accurate prediction of protein disorder segments with annotation for homology, secondary structure and linear motifs. Nucleic acids research 39:W190-W196
91. Wang C, Kurgan L (2018) Review and comparative assessment of similarity-based methods for prediction of drug-protein interactions in the druggable human proteome. Brief Bioinform
92. Wang C, Uversky VN, Kurgan L (2016) Disordered nucleiome: Abundance of intrinsic disorder in the DNA- and RNA-binding proteins in 1121 species from Eukaryota, Bacteria and Archaea. Proteomics 16:1486-1498
93. Ward JJ, Mcguffin LJ, Bryson K et al. (2004) The DISOPRED server for the prediction of protein disorder. Bioinformatics 20:2138-2139
94. Ward JJ, Sodhi JS, Mcguffin LJ et al. (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J Mol Biol 337:635-645
95. Wright PE, Dyson HJ (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. J Mol Biol 293:321-331
96. Wu Z, Hu G, Wang K et al. (2017) Exploratory Analysis of Quality Assessment of Putative Intrinsic Disorder in Proteins. In: 6th International Conference on Artificial Intelligence and Soft Computing. Zakopane, Poland, p 722-732
97. Wu Z, Hu G, Yang J et al. (2015) In various protein complexes, disordered protomers have large per-residue surface areas and area of protein-, DNA- and RNA-binding interfaces. FEBS Lett 589:2561-2569
98. Xie H, Vucetic S, Iakoucheva LM et al. (2007) Functional anthology of intrinsic disorder. 3. Ligands, post-translational modifications, and diseases associated with intrinsically disordered proteins. J Proteome Res 6:1917-1932
99. Xue B, Blocquel D, Habchi J et al. (2014) Structural disorder in viral proteins. Chem Rev 114:6880-6911
100. Xue B, Dunbrack RL, Williams RW et al. (2010) PONDR-FIT: a meta-predictor of intrinsically disordered amino acids. Biochim Biophys Acta 1804:996-1010
101. Xue B, Dunker AK, Uversky VN (2012) Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life. J Biomol Struct Dyn 30:137-149
102. Xue B, Mizianty MJ, Kurgan L et al. (2012) Protein intrinsic disorder as a flexible armor and a weapon of HIV-1. Cell Mol Life Sci 69:1211-1259
103. Xue B, Oldfield CJ, Dunker AK et al. (2009) CDF it all: consensus prediction of intrinsically disordered proteins based on various cumulative distribution functions. FEBS Lett 583:1469-1474

104. Xue B, Williams RW, Oldfield CJ et al. (2010) Archaic chaos: intrinsically disordered proteins in Archaea. BMC Syst Biol 4 Suppl 1:S1
105. Yan J, Dunker AK, Uversky VN et al. (2016) Molecular recognition features (MoRFs) in three domains of life. Molecular bioSystems 12:697-710
106. Yan J, Kurgan L (2017) DRNApred, fast sequence-based method that accurately predicts and discriminates DNA- and RNA-binding residues. Nucleic Acids Res 45:e84
107. Yan J, Marcus M, Kurgan L (2014) Comprehensively designed consensus of standalone secondary structure predictors improves Q3 by over 3%. J Biomol Struct Dyn 32:36-51
108. Yan J, Mizianty MJ, Filipow PL et al. (2013) RAPID: fast and accurate sequence-based prediction of intrinsic disorder content on proteomic scale. Biochim Biophys Acta 1834:1671-1680
109. Zhang J, Ma Z, Kurgan L (2017) Comprehensive review and empirical analysis of hallmarks of DNA-, RNA- and protein-binding residues in protein chains. Brief Bioinform
110. Zhang T, Faraggi E, Xue B et al. (2012) SPINE-D: accurate prediction of short and long disordered regions by a single neural-network based method. J Biomol Struct Dyn 29:799-813
111. Zhou JH, Zhao SW, Dunker AK (2018) Intrinsically Disordered Proteins Link Alternative Splicing and Post-translational Modifications to Complex Cell Signaling and Regulation. Journal of Molecular Biology 430:2342-2359