nature protocols

Review article

https://doi.org/10.1038/s41596-023-00876-x

Tutorial: a guide for the selection of fast and accurate computational tools for the prediction of intrinsic disorder in proteins

In the format provided by the authors and unedited

Evaluation metrics

Disorder predictors generate two types of outputs for each amino acid in a given protein sequence: a numeric propensity that quantifies likelihood that a given residue is disordered, and binary value that categorizes this residue as either disordered or structured. The predictive quality of propensities is typically evaluated with the Area Under receiver operating characteristic Curve (AUC) while the binary predictions are usually assessed with the Matthews Correlation Coefficient (MCC). For example, these metrics were utilized in the most recent community assessments[1,2]. MCC is defined as:

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}}$$

where TP (true positive) and TN (true negative) represent the number of correctly predicted disordered and structured amino acids, respectively, while FP (false positive) and FN (false negative) quantify the number of misclassified structured and disordered amino acids, respectively. MCC values range between -1 and 0, with -1 denoting inverted prediction (positives are predicted as negatives and vice versa), 0 for an inaccurate prediction, and 1 for highly accurate/perfect prediction.

The receiver operating characteristic (ROC)[3] curve is computed by thresholding the propensity values (i.e., thresholds are set to all distinct values of putative propensities) to obtain the corresponding binary predictions and computing TPR = TP/(TP+FN) and FPR = FP/(TN+FP) values for each threshold. The curve is composed of lines that connect consecutive (TPR, FPR) points and represents a relation between the true positive and false positive rates. AUC is the area under the ROC curve where higher values suggest that the predicted propensities are more accurate. AUC values range between 0.5 (equivalent to a random predictor where true positive and false positive rates are equal) and 1 (highly accurate/perfect prediction).

Supplementary Table S1. Intrinsic disorder predictors that are available to the end users. The methods are sorted in the chronological order of their year of publications. "Applies DNN" column identifies tools that use deep neural network (DNN) models. "Fast" column shows methods that predict disorder for a single protein in under 1 second; the runtime data was extracted from the CAID experiment[1] and a subsequent study[4]. "Availability" column explains how the predictors are distributed to the end user: as "SP" (standalone program) and/or "WS" (webserver).

Predictor name	Reference	Method number from Figure 1	Applies DNN	Fast	Availability	URL
DisEMBL-465	[5]	1	No	Yes	SP+WS	http://dis.embl.de/
DisEMBL-HL	[5]	2	No	Yes	SP+WS	http://dis.embl.de/
FoldUnfold	[6]	3	No	Yes	WS	http://bioinfo.protres.ru/ogu/
PONDR VSL2B	[7]	4	No		WS	http://www.pondr.com/
IsUnstruct	[8]	5	No	Yes	WS	http://bioinfo.protres.ru/IsUnstruct/
Espritz-DisProt	[9]	6	No		SP+WS	http://old.protein.bio.unipd.it/espritz/
Espritz-NMR	[9]	7	No		SP+WS	http://old.protein.bio.unipd.it/espritz/
Espritz-Xray	[9]	8	No		SP+WS	http://old.protein.bio.unipd.it/espritz/
DISOPRED3	[10]	9	No		SP+WS	http://bioinf.cs.ucl.ac.uk/psipred/
DisPredict	[11]	10	No		SP	https://github.com/tamjidul/DisPredict2_PSEE
MobiDB-lite	[12]	11	No		WS	http://mobidb.bio.unipd.it/
SPOT-Disorder	[13]	12	Yes		SP+WS	https://sparks-lab.org/server/spot-disorder/
IUpred2A-long	[14]	13	No	Yes	SP+WS	https://iupred2a.elte.hu/
IUpred2A-short	[14]	14	No	Yes	SP+WS	https://iupred2a.elte.hu/
pyHCA	Unpublished (released in 2018)	15	No		SP	https://github.com/T-B-F/pyHCA
SPOT-Disorder-Single	[15]	16	Yes		SP+WS	https://sparks-lab.org/server/spot-disorder-single/
rawMSA	[16]	17	Yes		SP	https://bitbucket.org/clami66/rawmsa/src/master/
SPOT-Disorder2	[17]	18	Yes		SP+WS	https://sparks-lab.org/server/spot-disorder2/
IDP-Seq2Seq	[18]	19	Yes		WS	http://bliulab.net/IDP-Seq2Seq/
flDPnn	[19]	20	Yes		SP+WS	http://biomine.cs.vcu.edu/servers/flDPnn/
Metapredict	[20]	21	Yes		SP+WS	https://github.com/idptools/metapredict
RFPR-IDP	[21]	22	Yes		WS	http://bliulab.net/RFPR-IDP/server
DisoMine	[22]	23	Yes		SP+WS	https://www.bio2byte.be/b2btools/disomine/

References

1 Necci, M., Piovesan, D., Predictors, C., DisProt, C. & Tosatto, S. C. E. Critical assessment of protein intrinsic disorder prediction. *Nat Methods* **18**, 472-481, doi:10.1038/s41592-021-01117-3 (2021).

2 Monastyrskyy, B., Kryshtafovych, A., Moult, J., Tramontano, A. & Fidelis, K. Assessment of protein disorder region predictions in CASP10. *Proteins* **82 Suppl 2**, 127-137, doi:10.1002/prot.24391 (2014).

3 Fawcett, T. An introduction to ROC analysis. *Pattern Recogn Lett* **27**, 861-874, doi:10.1016/j.patrec.2005.10.010 (2006).

4 Zhao, B. & Kurgan, L. Deep learning in prediction of intrinsic disorder in proteins. *Computational and Structural Biotechnology Journal* **20**, 1286-1294, doi:https://doi.org/10.1016/j.csbj.2022.03.003 (2022).

5 Linding, R. *et al.* Protein disorder prediction: implications for structural proteomics. *Structure* **11**, 1453-1459 (2003).

6 Galzitskaya, O. V., Garbuzynskiy, S. O. & Lobanov, M. Y. FoldUnfold: web server for the prediction of disordered regions in protein chain. *Bioinformatics* **22**, 2948-2949, doi:10.1093/bioinformatics/btl504 (2006).

7 Peng, K., Radivojac, P., Vucetic, S., Dunker, A. K. & Obradovic, Z. Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics* **7**, 208, doi:10.1186/1471-2105-7-208 (2006).

8 Lobanov, M. Y. & Galzitskaya, O. V. The Ising model for prediction of disordered residues from protein sequence alone. *Phys Biol* **8**, 035004, doi:10.1088/1478-3975/8/3/035004 (2011).

9 Walsh, I., Martin, A. J., Di Domenico, T. & Tosatto, S. C. ESpritz: accurate and fast prediction of protein disorder. *Bioinformatics* **28**, 503-509, doi:10.1093/bioinformatics/btr682 (2012).

10 Jones, D. T. & Cozzetto, D. DISOPRED3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics* **31**, 857-863, doi:10.1093/bioinformatics/btu744 (2015).

11 Iqbal, S. & Hoque, M. T. DisPredict: A Predictor of Disordered Protein Using Optimized RBF Kernel. *PloS one* **10**, e0141551, doi:10.1371/journal.pone.0141551 (2015).

12 Necci, M., Piovesan, D., Dosztanyi, Z. & Tosatto, S. C. E. MobiDB-lite: fast and highly specific consensus prediction of intrinsic disorder in proteins. *Bioinformatics* **33**, 1402-1404, doi:10.1093/bioinformatics/btx015 (2017).

13 Hanson, J., Yang, Y., Paliwal, K. & Zhou, Y. Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks. *Bioinformatics* **33**, 685-692, doi:10.1093/bioinformatics/btw678 (2017).

14 Meszaros, B., Erdos, G. & Dosztanyi, Z. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res* **46**, W329-W337, doi:10.1093/nar/gky384 (2018).

15 Hanson, J., Paliwal, K. & Zhou, Y. Accurate Single-Sequence Prediction of Protein Intrinsic Disorder by an Ensemble of Deep Recurrent and Convolutional Architectures. *J Chem Inf Model* **58**, 2369-2376, doi:10.1021/acs.jcim.8b00636 (2018).

16 Mirabello, C. & Wallner, B. rawMSA: End-to-end Deep Learning using raw Multiple Sequence Alignments. *PloS one* **14**, e0220182, doi:10.1371/journal.pone.0220182 (2019).

17 Hanson, J., Paliwal, K. K., Litfin, T. & Zhou, Y. SPOT-Disorder2: Improved Protein Intrinsic Disorder Prediction by Ensembled Deep Learning. *Genomics Proteomics Bioinformatics* **17**, 645-656, doi:10.1016/j.gpb.2019.01.004 (2019).

18 Tang, Y. J., Pang, Y. H. & Liu, B. IDP-Seq2Seq: identification of intrinsically disordered regions based on sequence to sequence learning. *Bioinformatics* **36**, 5177-5186, doi:10.1093/bioinformatics/btaa667 (2021).

19 Hu, G. *et al.* flDPnn: Accurate intrinsic disorder prediction with putative propensities of disorder functions. *Nat Commun* **12**, 4438, doi:10.1038/s41467-021-24773-7 (2021).

20 Emenecker, R. J., Griffith, D. & Holehouse, A. S. Metapredict: a fast, accurate, and easy-to-use predictor of consensus disorder and structure. *Biophys J* **120**, 4312-4319, doi:10.1016/j.bpj.2021.08.039 (2021).

21 Liu, Y., Wang, X. & Liu, B. RFPR-IDP: reduce the false positive rates for intrinsically disordered protein and region prediction by incorporating both fully ordered proteins and disordered proteins. *Briefings in bioinformatics* **22**, 2000-2011, doi:10.1093/bib/bbaa018 (2021).

22 Orlando, G., Raimondi, D., Codice, F., Tabaro, F. & Vranken, W. Prediction of Disordered Regions in Proteins with Recurrent Neural Networks and Protein Dynamics. *J Mol Biol* **434**, 167579, doi:10.1016/j.jmb.2022.167579 (2022).