Exploratory Analysis of Quality Assessment of Putative Intrinsic Disorder in Proteins

Zhonghua Wu[1], Gang Hu[1], Kui Wang[1], and Lukasz Kurgan[2]

[1] School of Mathematical Sciences and LPMC, Nankai University, Tianjin, People's Republic of China
{wuzhh, huggs, wangkui}@nankai.edu.cn
[2] Department of Computer Science, Virginia Commonwealth University, Richmond, VA, USA
lkurgan@vcu.edu

Abstract. Intrinsically disorder proteins are abundant in nature and can be accurately identified from sequences using computational predictors. While predictions of disorder are relatively easy to obtain there are no tools to assess their quality for a particular amino acid or protein. Quality assessment (QA) scores that quantify correctness of the predictions are not available. We define QA for the prediction of intrinsic disorder and use a large dataset of over 25 thousand proteins and ten modern predictors of disorder to empirically assess the first approach to quantify QA scores. We formulate the QA scores based on the readily available propensities of the intrinsic disorder generated by the ten methods. Our evaluation reveals that these QA scores offer good predictive performance for native structured residues (AUC>0.74) and poor predictive performance for native disordered residues (AUC<0.67). Specifically, we show that most of the native disordered residues that are incorrectly predicted as structured have high QA values that inaccurately suggest that these predictions are correct. Consequently, more research is needed to develop high-quality QA scores. We also outline three possible future research directions.

1 Introduction

Intrinsically disordered proteins lack stable tertiary structure under physiological conditions along their entire amino acid chain or in specific region(s) [1, 2]. They are abundant in nature, with recent estimates showing that about 19% of amino acids in eukaryotic proteins are disordered [3], and up to 50% of eukaryotic proteins have at least one long (≥ 30 consecutive amino acids) intrinsically disordered region [4, 5]. Intrinsically disordered proteins are crucial for a diverse range of cellular functions including transcription, translation, signaling, protein-protein, protein-nucleic acids and virus-host interactions, to name just a few [2, 3, 6, 7]. A large number of computational methods that predict intrinsic disorder in protein sequences was developed. A study from 2012 estimates this number to be at about 60 [8]. The predictions that these methods generate are utilized to support and plan experimental studies and to quantify prevalence and analyze functions of disorder on a large, genomic scale [3, 9-13]. They are also used in other research areas including structural genomics [14]. In

© Springer-Verlag Berlin Heidelberg 2011

recent years two large databases that offer access to putative annotations of intrinsic disorder for millions of proteins were developed: MobiDB [15, 16] and D^2P^2 [17].

In spite of the popularity and wide-spread use of these predictors and the fact that their predictive performance was evaluated in a number of studies [18-23], there are no studies that investigate quality assessment of these predictions. While the users nowadays can easily collect predictions of disorder, there are no methods that quantify quality of these predictions for a particular amino acid or protein. In other words, quality assessment methods that assign a numeric score to each prediction that quantifies whether it is correct are lacking. This is in stark contrast to the prediction of the tertiary protein structure where many tools for the quality assessment were developed in recent years [24-28]. To this end, we define the quality assessment in the context of the prediction of intrinsic disorder. Using a large dataset of proteins, we also empirically assess whether the propensities of the intrinsic disorder generated by the modern predictors of disorder can be used as a proxy for the quality assessment scores.

2 Materials and Methods

2.1 Dataset

The dataset was originally developed in ref. [22] and can be downloaded from http://mobidb.bio.unipd.it/lsd. Proteins from the UniProt resource [29] were mapped into the MobiDB database [30] to obtain their annotations of native intrinsic disorder. All proteins for which the annotations were found are included and a majority vote was used to assign disorder in cases when multiple annotations are found in MobiDB, i.e., a given residue is assumed to be disordered if most of the annotations for this residues indicate that it is disordered. This approach arguably allows to filter out conflicts due to variations in experimental conditions [22]. Similar sequences were removed at 90% pairwise sequence identity using the CD-HIT program [31] resulting in a set of 25,883 annotated proteins. Each residue in these proteins is annotated as disordered, structured or unknown, the latter in the case when MobiDB does not provide an annotation. Our analysis is based on the residues that are annotated as either disordered or structured. We exclude the residues with the unknown annotations. Moreover, the dataset is further reduced to 25,717 proteins for which we were able to secure putative intrinsic disorder with all considered predictors of disorder. In total, our dataset includes 7,049,517 annotated residues with 6,700,101 and 349,416 that are structured and disordered, respectively. This corresponds to the overall disorder content (defined as a fraction of disordered residues among all residues) of 5%.

2.2 Putative annotations of intrinsic disorder

Putative annotations of intrinsic disorder were generated with ten methods: three version of the ESpritz method [32] that are designed to predict disorder annotated using X-ray crystallography (Espritz-X-ray), NMR (Espritz-NMR), and DisProt database [33] (Espritz-Disprot); two versions of IUPred [34] that are optimized to predict short (IUPred-short) and long (IUPred-long) disordered regions; two versions of the

DisEMBL method [35] that predict disordered regions defined as hot loops (DisEMBL-HL) and based on the remark 465 from Protein Data Bank (PDB) [36] (DisEMBL-465), RONN [37],VSL2b [38], and GlobPlot that predicts globular regions [39]. These methods represent a comprehensive selection of modern predictors that cover various flavors of disorder and that are sufficiently runtime-efficient to provide results at the scale of our large dataset; the runtime of these methods is under 1 minute for an average size protein sequence. The predictors of disorder typically generate two outputs for each residue in the input protein sequence: a real-valued propensity score and a binary prediction. The score is a putative likelihood that a given residue is in a disordered conformation. The binary value is usually derived from the propensity based on a method-specific threshold and it categorizes the residue as either disordered or structured. Residue with propensities > threshold are classified as disordered and the remaining residues are classified as structured. The ranges of values of the propensities for the ten predictors together with the native and putative disorder content, the latter estimated from the putative binary values, are summarized in Table 1. Interestingly, the Pearson correlation coefficients (PCCs) between predicted propensities generated by different predictors range between 0.07 and 0.81, with average of 0.46. This demonstrates that these methods in fact offer substantially different predictions.

Table 1. The native and predicted amount of intrinsic disorder for the benchmark set of 25717 proteins. We also list the minimal and maximal values of propensity and the threshold value used to convert these propensities into binary scores for the 10 preditors of intrinsic disorder.

Native annotations and predictors	**Putative propensity of disorder**			**Disorder content**
	min	**max**	**threshold**	
Native annotations	NA	NA	NA	5.0%
DisEMBL-465	0.000	0.968	0.500	6.4%
DisEMBL-HL	0.000	0.585	0.086	28.9%
Espritz-Disprot	0.004	0.978	0.507	2.6%
Espritz-NMR	0.002	0.997	0.309	9.1%
Espritz-X-ray	0.003	0.997	0.143	16.5%
GlobPlot	-0.329	0.513	0.000	13.5%
IUPred-long	0.000	0.995	0.500	6.0%
IUPred-short	0.000	1.000	0.500	6.7%
RONN	0.070	1.000	0.500	16.2%
VSL2b	0.002	1.000	0.500	21.0%

2.3 Definition of quality assessment for putative intrinsic disorder

The putative annotations of intrinsic disorder are typically derived based on the binary values where residue are categorized as either structured or disordered. The putative propensities can be used to quantify confidence that accompanies the binary predictions. The putative disordered residues predicted with high propensity scores should be more accurately predicted compared to the residues that are associated with propensities that are just slightly higher than the threshold. The same is true for the structured residues where the putative structured residues that have low propensities should be more accurately predicted than the structured residues with propensities just

below the threshold. However, while predictive performance of the disorder predictors was evaluated extensively [18-23], the use of the propensities as a proxy to quantify quality of these predictions was not yet researched.

The quality assessment (QA) boils down to computation of a score that quantifies correctness of a given prediction. More specifically, in the QA scenario each prediction, whether it suggests that a given residue is disordered or structured, is associated with a propensity score that is high when the prediction is correct and low when it is incorrect. In other words, native disordered residues predicted as disordered and native structured residues predicted as structured should have high QA scores, while residues that are incorrectly predicted (native disordered as structured or native structured as disordered) should have low QA scores. One immediately available option to generate these QA scores is to use the predicted propensities for disorder to generate QA scores for the binary disorder predictions:

IF $D_{prop} > THR$ THEN $QA_{score} = \{(D_{prop} - THR) / (\max(D_{prop}) - THR)\}$

IF $D_{prop} \leq THR$ THEN $QA_{score} = \{(D_{prop} - THR) / (\min(D_{prop}) - THR)\}$

where D_{prop} is the putative propensity for disorder and THR is the threshold used to convert D_{prop} into the binary disorder prediction. This definition ensures that high and low values of the putative propensity for disorder (that denote likely correct predictions of disordered and structured residues, respectively) correspond to high QA scores, while QA scores for values of the predicted propensity for disorder that are close to the threshold are low. The relation between values of D_{prop} and QA_{score} is visualized in Fig. 1.

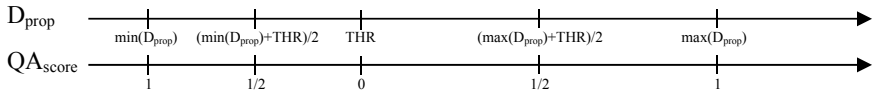


Fig. 1. The relation between the values of the putative propensity for intrinsic disorder (D_{prop}) and the values of the quality assessment score (QA_{score}).

2.4 Evaluation measures

Quality of the predicted propensity for the intrinsic disorder is typically evaluated using ROC curves and the corresponding AUC values [18-23]. More specifically, the propensities are used to compute a curve defined by FPR = FP / (FP+TN) and TPR = TP / (TP+FN) values where TP is the number of correctly predicted disordered residues, FP is the number of structured residues predicted as disordered, TN is the number of correctly predicted structured residues, and FN is the number of disordered residues predicted as structured; multiple values of FPR and TPR are generated by using different thresholds on the value of the propensity. AUC is the area under the ROC curve and its values range between 0.5 for a random-like prediction and 1 for the perfect prediction. We denote this measure as AUCd (AUC for the prediction of disordered residues).

We similarly utilize the AUC values to assess the predictive quality of the QA scores. In this case TP is the number of correctly predicted correct predictions (of both disordered and structured residues) based on the QA scores, FP is the number

incorrect predictions that are predicted as correct using QA scores, TN is the number of correctly predicted incorrect predictions (of both disordered and structured residues) and FN is the number of correct predictions that are predicted as incorrect utilizing the QA scores. We denote this measure as AUCqa (AUC for the quality assessment). We also compute the AUCqa values specifically for the native disordered residues (AUCqa_d) and for the native structured residues (AUCqa_s). The latter two values quantify how well the quality assessment scores work when they are used for the native disordered and native structured residues.

3 Results

The AUCd values of the considered ten predictors of the intrinsic disorder are in agreement with the results in [22]. The values are shown in Fig. 2 and they range between 0.63 and 0.81 with average of 0.75. These results are also similar to the findings in [21] where the AUCd values of 19 predictors are shown to be between 0.70 and 0.82. Collectively, these studies conclude that the predictors of intrinsic disorder offer relatively strong predictive performance. However, the binary predictions of some of these methods disagree with the native annotations of disorder. Table 1 reveals that the native disorder content in our large dataset is at 5% while the putative disorder content generated by the ten predictors varies between 2.6% and 28.9%, with an average of 12.7%. This suggests that the putative binary annotation require improvements, and this could be addressed by coupling them with the QA scores.

Fig. 2 summarizes the AUCqa values for the QA scores that were computed from the putative propensities for disorder. These AUCqa values quantify how well the QA scores predict correctness of the binary predictions of disorder. The size of the circles represent relative values of the AUCqa and the absolute values are shown next to the circles. The AUCqa values range between 0.74 for VSL2b and 0.90 for Espritz-X-ray, with an average value of 0.81. Fig. 2 also shows the AUCqa_d and AUCqa_s values (the AUC for the QA scores for the native disordered and structured residues, respectively) as the *y*- and *x*-axis coordinates, respectively. The bubbles located below the 0.5 value on the *y*-axis correspond to seven methods that perform very poorly for the disordered residues: IUPred-short, IUPred-long, Espritz-NMR, DisEMBL-HL, GlobPlot, Espritz-X-ray, and Espritz-Disprot. While their overall AUCqa values are relatively high (between 0.77 and 0.90), they provide high quality QA scores only for the structured residues. In other words, the QA scores for these seven methods successfully identify correctly vs. incorrectly predicted structured residues, while they largely fail to identify correctly predicted disordered residues. Their AUCqa values are high in spite of the low values of AUCqa_d because a significant majority of the residues in the dataset is structured. Two methods, DisEMBL-465 and RONN, achieve modest values of AUCqa <0.8 coupled with relatively high AUCqa_s at about 0.82 and slightly above-random values of AUCqa_d at about 0.55. The only method for which the QA scores are reasonably balanced between the structured and disordered residues is VSL2b. It secures AUCqa = 0.74, AUCqa_d = 0.67 and AUCqa_s =

0.74. However, these are rather modest values of predictive performance, particularly the AUCqa_d for the QA scores for the intrinsically disordered residues.

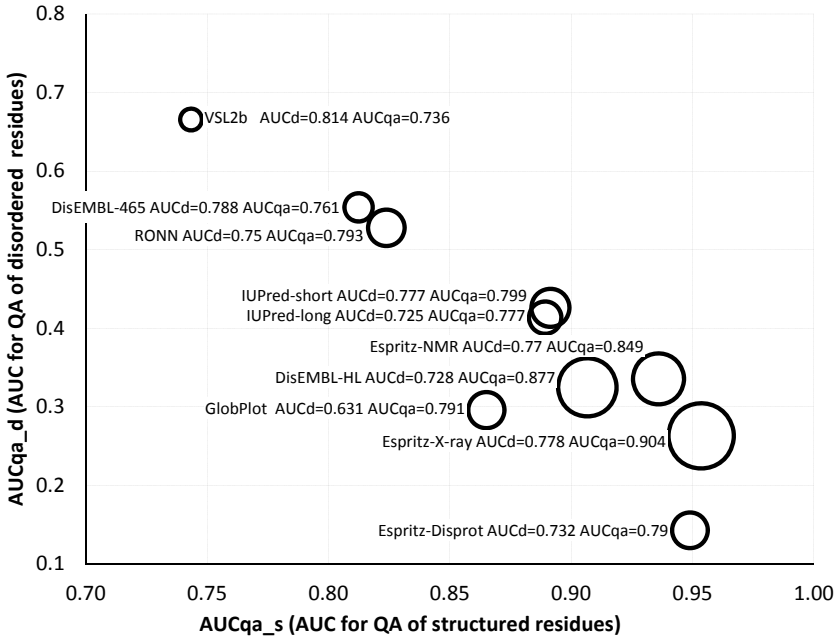
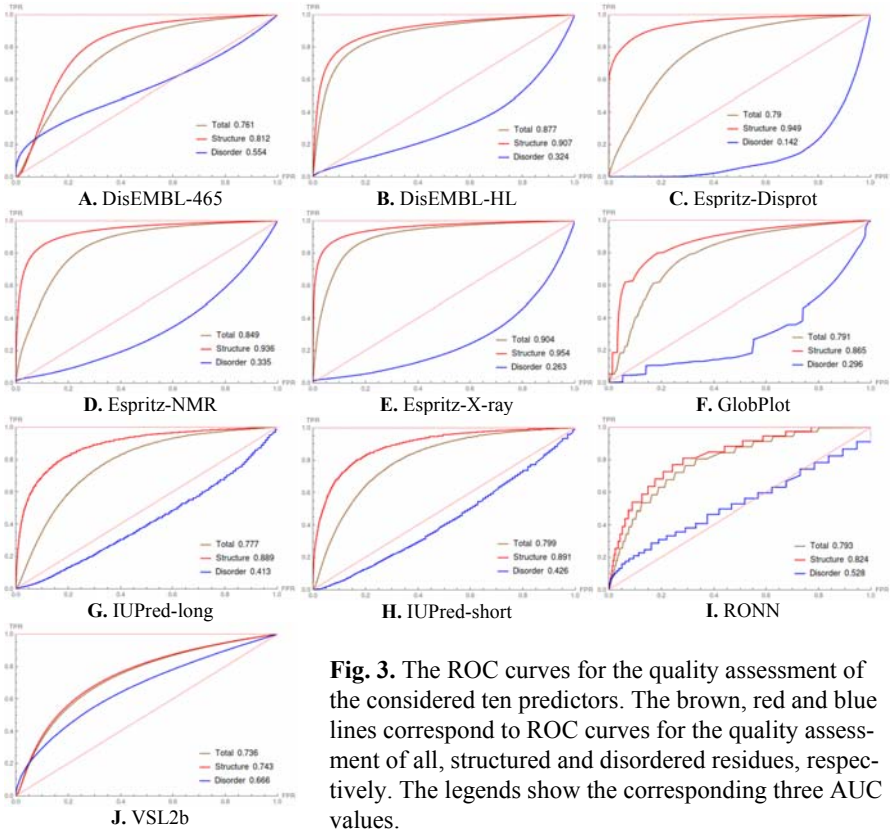


Fig. 2. Relation between AUC for the quality assessment of the disordered residues (AUCqa_d on the *y*-axis) and structured residues (AUCqa_s on the *x*-axis). Each predictor is represented by a circle; sizes of the circles represent relative values of AUC of the quality assessment of all residues (AUCqa). The names of the predictors together with the numeric values of AUCqa and AUCd (AUC for the prediction of disordered residues) are shown next to the circles.



A. DisEMBL-465 **B.** DisEMBL-HL **C.** Espritz-Disprot

D. Espritz-NMR **E.** Espritz-X-ray **F.** GlobPlot

G. IUPred-long **H.** IUPred-short **I.** RONN

J. VSL2b

Fig. 3. The ROC curves for the quality assessment of the considered ten predictors. The brown, red and blue lines correspond to ROC curves for the quality assessment of all, structured and disordered residues, respectively. The legends show the corresponding three AUC values.

Fig. 3 that gives the ROC curves for the QA scores offers further insights. The brown and red lines that denote the ROC curves for the quality assessment of all and native structured residues, respectively, reveal a favorable trade-off between the TPR and FPR values, i.e., TPR values are substantially higher than the corresponding FPR values. The three exceptions include DisEMBL-465, RONN and VSL2b methods (Figs 3A, 3I and 3J) for which the two curves are relatively flat for the low FPR values, resulting in low AUCqa and AUCqa_s values. More importantly, the blue ROC curves for the quality assessment of native disordered residues, which for most predictors are located below the diagonal line, demonstrate that the corresponding FPR values are higher than the TPR values. Consequently, the QA scores produce more incorrect predictions than the number of correct predictions. More specifically, the ratios of incorrect predictions of disordered residues that are predicted as correct using the QA scores among all incorrect predictions (FPR values) are higher than the ratios of correct predictions of disordered residues that are predicted as correct using the QA scores among all correct predictions (TPR values). In other words, the high FPR means that many native disordered residues that are incorrectly predicted as

structured are associated with high QA values. Such high QA values inaccurately suggest that the associated with them predictions are correct. In turn, the high QA values result from the fact that the corresponding putative propensities for disorder are low for these disordered residues. We observe that virtually all of the considered methods, except for VSL2b, generate low putative propensities for a majority of the disordered residues. This is a significant drawback of the putative propensities for disorder generated by the considered representative set of disorder predictors. It effectively renders the corresponding QA scores useless when applied to the native disordered residues. Finally, although the three ROC curves for the QA scores for VLS2b are all above the diagonal line (Fig. 3J), the curves are flat suggesting that the predictive performance of these scores is rather low.

4 Conclusions

Our analysis that examines ten modern predictors of intrinsic disorder on a large set of close to 26 thousand proteins reveals that although the overall predictive performance of these methods is relatively high, the putative annotations that they generate would benefit from the inclusion of QA scores. These scores could be used to indicate which predictions could be trusted more than others and to identify correct vs. incorrect predictions. We are the first to attempt to define the QA scores based on the readily available putative propensities for disorder generated by the ten predictors. We empirically evaluate whether these propensities can be used to derive accurate QA values for the assessment of the corresponding binary predictions of disorder. Our analysis demonstrates that the QA scores that we define provide accurate results for the native structured residues for majority of the considered methods. However, the QA scores for the native disordered residues are inaccurate. For 9 out of 10 methods most of the native disordered residues that are incorrectly predicted as structured have high QA values (and low putative propensities for disorder) which falsely indicate that the corresponding predictions are correct. The only method for which QA scores perform reasonably well for both structured and disordered residues is VSL2b. However, the predictive quality of its QA scores is relatively modest, with the AUC values equal 0.74 and 0.67 for the structured and disordered residues, respectively.

Our results suggest that the QA scores generated based on the propensity for intrinsic disorder generated by modern, high-throughput predictors do not offer desirable levels of predictive performance. Further research to develop high-quality QA scores for the putative intrinsic disorder is needed. This is particularly urgent in the context of the recent emergence of large databases, such as MobiDB and D^2P^2, which offer easy access to predicted disorder for dozens of millions of proteins. Three possible directions could be pursued. The first option is to build one methodology that will provide QA scores for any disorder predictor using predictions from a single method. This would be very challenging given the relatively high degree of differences between the predictions from different methods for the same protein sequence. The second alternative is to develop one methodology that will provide QA scores for any disorder predictor using predictions from multiple methods. In other words, predic-

tions from several disorder predictors would be used to derive a generic QA score which could be used to assess predictions of any of the input methods. While this should be easier than the first alternative, it will also require availability of multiple disorder predictions. This is feasible when employing the MobiDB and D^2P^2 databases that provides access to multiple predictions for each protein. The third option is to build QA methodologies that are coupled with specific disorder predictors. This option would be perhaps the easiest to develop but it would also require designing multiple QA schemes, as many as the number of the corresponding disorder predictors.

Acknowledgments

We thank Dr. Silvio Tosatto and his research group from University of Padova for have sharing their dataset and predictions of disorder, which they published in ref. [22]. This research was supported in part by the National Science Foundation grant 1617369 and by the Qimonda Endowed Chair from Virginia Commonwealth University to L.K.

References

1. Dunker, A.K., Babu, M.M., Barbar, E., Blackledge, M., Bondos, S.E., Dosztányi, Z., Dyson, H.J., Forman-Kay, J., Fuxreiter, M., Gsponer, J., Han, K.-H., Jones, D.T., Longhi, S., Metallo, S.J., Nishikawa, K., Nussinov, R., Obradovic, Z., Pappu, R.V., Rost, B., Selenko, P., Subramaniam, V., Sussman, J.L., Tompa, P., Uversky, V.N.: What's in a name? Why these proteins are intrinsically disordered. Intrinsically Disordered Proteins 1, e24157 (2013)
2. 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.: Classification of Intrinsically Disordered Regions and Proteins. Chemical Reviews 114, 6589-6631 (2014)
3. Peng, Z., Yan, J., Fan, X., Mizianty, M.J., Xue, B., Wang, K., Hu, G., Uversky, V.N., Kurgan, L.: Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life. Cell Mol Life Sci 72, 137-151 (2015)
4. Xue, B., Dunker, A.K., Uversky, V.N.: 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 (2012)
5. Ward, J.J., Sodhi, J.S., McGuffin, L.J., Buxton, B.F., Jones, D.T.: Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J Mol Biol 337, 635-645 (2004)
6. Fuxreiter, M., Toth-Petroczy, A., Kraut, D.A., Matouschek, A., Lim, R.Y., Xue, B., Kurgan, L., Uversky, V.N.: Disordered proteinaceous machines. Chem Rev 114, 6806-6843 (2014)
7. Xue, B., Blocquel, D., Habchi, J., Uversky, A.V., Kurgan, L., Uversky, V.N., Longhi, S.: Structural disorder in viral proteins. Chem Rev 114, 6880-6911 (2014)

8. Kozlowski, L.P., Bujnicki, J.M.: MetaDisorder: a meta-server for the prediction of intrinsic disorder in proteins. BMC Bioinformatics 13, 1-11 (2012)
9. Peng, Z., Oldfield, C.J., Xue, B., Mizianty, M.J., Dunker, A.K., Kurgan, L., Uversky, V.N.: A creature with a hundred waggly tails: intrinsically disordered proteins in the ribosome. Cell Mol Life Sci 71, 1477-1504 (2014)
10. Xue, B., Mizianty, M.J., Kurgan, L., Uversky, V.N.: Protein intrinsic disorder as a flexible armor and a weapon of HIV-1. Cell Mol Life Sci 69, 1211-1259 (2012)
11. Pentony, M.M., Jones, D.T.: Modularity of intrinsic disorder in the human proteome. Proteins 78, 212-221 (2010)
12. Wang, C., Uversky, V.N., Kurgan, L.: 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 (2016)
13. Peng, Z., Xue, B., Kurgan, L., Uversky, V.N.: Resilience of death: intrinsic disorder in proteins involved in the programmed cell death. Cell Death Differ 20, 1257-1267 (2013)
14. Oldfield, C.J., Xue, B., Van, Y.Y., Ulrich, E.L., Markley, J.L., Dunker, A.K., Uversky, V.N.: Utilization of protein intrinsic disorder knowledge in structural proteomics. Biochim Biophys Acta 1834, 487-498 (2013)
15. Potenza, E., Domenico, T.D., Walsh, I., Tosatto, S.C.E.: MobiDB 2.0: an improved database of intrinsically disordered and mobile proteins. Nucleic Acids Research 43, D315-D320 (2015)
16. Di Domenico, T., Walsh, I., Martin, A.J.M., Tosatto, S.C.E.: MobiDB: a comprehensive database of intrinsic protein disorder annotations. Bioinformatics 28, 2080-2081 (2012)
17. Oates, M.E., Romero, P., Ishida, T., Ghalwash, M., Mizianty, M.J., Xue, B., Dosztányi, Z., Uversky, V.N., Obradovic, Z., Kurgan, L., Dunker, A.K., Gough, J.: D2P2: database of disordered protein predictions. Nucleic Acids Research 41, D508-D516 (2013)
18. Deng, X., Eickholt, J., Cheng, J.: A comprehensive overview of computational protein disorder prediction methods. Mol Biosyst 8, 114-121 (2012)
19. Monastyrskyy, B., Fidelis, K., Moult, J., Tramontano, A., Kryshtafovych, A.: Evaluation of disorder predictions in CASP9. Proteins 79 Suppl 10, 107-118 (2011)
20. Monastyrskyy, B., Kryshtafovych, A., Moult, J., Tramontano, A., Fidelis, K.: Assessment of protein disorder region predictions in CASP10. Proteins 82 Suppl 2, 127-137 (2014)
21. Peng, Z.L., Kurgan, L.: Comprehensive comparative assessment of in-silico predictors of disordered regions. Curr Protein Pept Sci 13, 6-18 (2012)
22. Walsh, I., Giollo, M., Di Domenico, T., Ferrari, C., Zimmermann, O., Tosatto, S.C.: Comprehensive large-scale assessment of intrinsic protein disorder. Bioinformatics 31, 201-208 (2015)
23. Noivirt-Brik, O., Prilusky, J., Sussman, J.L.: Assessment of disorder predictions in CASP8. Proteins 77 Suppl 9, 210-216 (2009)
24. Kihara, D., Chen, H., Yang, Y.D.: Quality assessment of protein structure models. Curr Protein Pept Sci 10, 216-228 (2009)
25. Skwark, M.J., Elofsson, A.: PconsD: ultra rapid, accurate model quality assessment for protein structure prediction. Bioinformatics 29, 1817-1818 (2013)

26. McGuffin, L.J., Buenavista, M.T., Roche, D.B.: The ModFOLD4 server for the quality assessment of 3D protein models. Nucleic Acids Res 41, W368-372 (2013)
27. Cao, R., Bhattacharya, D., Adhikari, B., Li, J., Cheng, J.: Massive integration of diverse protein quality assessment methods to improve template based modeling in CASP11. Proteins 84 Suppl 1, 247-259 (2016)
28. Kryshtafovych, A., Fidelis, K.: Protein structure prediction and model quality assessment. Drug Discov Today 14, 386-393 (2009)
29. UniProt, C.: UniProt: a hub for protein information. Nucleic Acids Res 43, D204-212 (2015)
30. Potenza, E., Di Domenico, T., Walsh, I., Tosatto, S.C.: MobiDB 2.0: an improved database of intrinsically disordered and mobile proteins. Nucleic Acids Res 43, D315-320 (2015)
31. Fu, L., Niu, B., Zhu, Z., Wu, S., Li, W.: CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28, 3150-3152 (2012)
32. Walsh, I., Martin, A.J., Di Domenico, T., Tosatto, S.C.: ESpritz: accurate and fast prediction of protein disorder. Bioinformatics 28, 503-509 (2012)
33. Sickmeier, M., Hamilton, J.A., LeGall, T., Vacic, V., Cortese, M.S., Tantos, A., Szabo, B., Tompa, P., Chen, J., Uversky, V.N., Obradovic, Z., Dunker, A.K.: DisProt: the Database of Disordered Proteins. Nucleic Acids Res 35, D786-793 (2007)
34. Dosztanyi, Z., Csizmok, V., Tompa, P., Simon, I.: IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics 21, 3433-3434 (2005)
35. Linding, R., Jensen, L.J., Diella, F., Bork, P., Gibson, T.J., Russell, R.B.: Protein disorder prediction: implications for structural proteomics. Structure 11, 1453-1459 (2003)
36. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E.: The Protein Data Bank. Nucleic Acids Res 28, 235-242 (2000)
37. Yang, Z.R., Thomson, R., McNeil, P., Esnouf, R.M.: RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. Bioinformatics 21, 3369-3376 (2005)
38. Peng, K., Radivojac, P., Vucetic, S., Dunker, A.K., Obradovic, Z.: Length-dependent prediction of protein intrinsic disorder. BMC Bioinformatics 7, 208 (2006)
39. Linding, R., Russell, R.B., Neduva, V., Gibson, T.J.: GlobPlot: Exploring protein sequences for globularity and disorder. Nucleic Acids Res 31, 3701-3708 (2003)