Databases of Protein Structure and Function Predictions at the Amino Acid Level

Bi Zhao[1], Lukasz Kurgan[1*]

[1]Department of Computer Science, Virginia Commonwealth University, Richmond, Virginia, United States

*Corresponding author: Lukasz Kurgan (lkurgan@vcu.edu)

Abstract

The rapid growth of the number of protein sequences greatly exceeds the pace of efforts to functionally and structurally annotate these proteins. The closing of the ensuing large and growing gap in the amino acid (AA)-level annotations of protein structure and function can be facilitated using accurate and fast computational predictors. Hundreds of sequence-based predictors of the AA-level annotations have been developed, making it challenging for the end users to identify suitable/good predictors and collect their results. One convenient solution is to obtain pre-computed predictions from large-scale databases, which include MobiDB, D^2P^2 and DescribePROT. These databases provide access to a diverse set of structural and functional characteristics, such as domains, secondary structures, solvent accessibility, intrinsic disorder, posttranslational modifications (PTMs), protein/DNA/RNA-binding AAs, disordered linkers and signal peptides. We motivate and introduce these databases, discuss and compare their contents, and comment on their applications and limitations. We find that these databases provide complementary scope and services, with D^2P^2 delivering comprehensive annotations of domains and PTMs, MobiDB focusing on the intrinsic disorder and being highly-connected to other resources, and DescribePROT covering the most diverse set of structural and functional features. We briefly examine practical applications for some of the structural predictions covered by these databases. We also concisely discuss modern predictive webservers that can be used when users need to collect the AA-level annotations for proteins that are not included in these databases.

1 Introduction

We face an enormous challenge to functionally and structurally characterize hundreds of millions of protein sequences [1, 2]. The current 2021_04 version UniProt includes 225.01 million of proteins and has more than tripled in size compared to the version 2016_04 from just 5 years ago that featured 63.69 million proteins [2, 3]. These annotations are done at three levels: atomic, amino acid (AA) and whole protein. The arguably most popular atomic-level database, Protein Data Bank (PDB) [4], covers 185 thousand protein structures. The most popular protein-level database, UniProt, has 565 thousand manually curated proteins (Swiss-Prot) and close to 225 million proteins with alignment-generated/predicted annotations (TrEMBL) [2]. The AA-level annotations bridge the gap between the atomic and protein-level annotations. They are computed from the PDB files and extracted from a sparsely populated

subset of the UniProt records. However, only a small fraction of AAs was annotated so far. Computational methods that predict the AA-level annotations from protein sequences (i.e., sequence-based predictors), many of which are described in this book, are widely used to assist with closing the huge and rapidly growing gap in the AA-level annotations.

The sequence-based predictors output AA-level annotations using predictive models trained and validated/tested using the ground truth generated by experimental methods, typically collected from PDB or related/derived databases, such as BioLip [5] or DisProt [6]. They often rely on models produced by machine learning (ML) algorithms. ML algorithms utilize experimentally annotated training datasets to parametrize models to "optimally" differentiate between AAs that have a given function/structure and the remaining non-functional/non-structural AAs. The training sets are two orders of magnitude larger than the corresponding set of training proteins since they concern AAs; average protein sequence has around 300 AAs [7]. Consequently, the amount of the experimentally annotated training data is sufficient to train and test accurate predictive models using sophisticated ML algorithms, such as deep neural networks. We stress that these models are optimized to provide accurate predictions for proteins that share low levels of similarity/homology with the proteins in the training dataset, typically < 30% similarity. In essence, the sequence-based *ab initio* methods can be used to make AA-level predictions for any of the 225 million of the sequenced proteins.

Hundreds of the sequence-based predictors of the AA-level annotations have been developed. They can be divided into two major groups: (1) methods that target prediction of functional AAs; and (2) methods that predict structural characteristics of AAs. The first group covers a broad spectrum of functions including prediction of AAs that interact with RNA, DNA, lipids and proteins, catalytic residues, cleavage and post-translational modification sites (PTMs), and intrinsic disorder. Selected, popular examples include DP-Bind [8, 9] and DBS-PSSM [10] that predict DNA-binding AAs; RNABindR [11-13] and Pprint [14] that identify putative RNA-binding residues; BindN+ [15], DRNApred [16] and NucBind [17] that predict DNA and RNA binding AAs; SPPIDER [18], PSIVER [19] and SCRIBER [20] that find putative protein-binding residues; DisoLipPred [21] that predicts lipid-binding AAs; PROSPERous [22] and DeepCleave [23] that generate putative cleavage sites; INTREPID [24, 25], PREvaIL [26] and CRpred [27] that produce putative catalytic residues; NetPhosK [28], SUMOsp [29, 30] and UbPred [31] that find putative PTMs; SignalP [32-35] and ChloroP [36] that identify putative signal peptides; IUPred [37-40], DISOPRED [41, 42] and flDPnn [43] that predict intrinsic disorder; and DisoRDPbind [44-46] and DeepDISObind [47] that generate putative disordered residues that interact with DNA, RNA and proteins. The second category targets prediction of various structural features of the AAs including their secondary structure, torsion angles, solvent accessibility, flexibility and residue-residue contacts. Example popular predictors include PSIPRED [48, 49], PHD [50, 51] and JPRED [52-54] that predict secondary structure; PHDacc [55] and ACCpro [56] that predict solvent accessibility; PROFbval [57, 58] and FlexRP [59] that generate putative flexibility; and PSICOV [60], GREMLIN [61], ContactMap [62] and SVMcon [63] that produce putative residue-residue contacts. There are many more methods that target prediction of each of these structural and functional characteristics. For instance, there are over 100 predictors of intrinsic disorder [64-67], over 60 tools for the prediction of secondary structure [68-71], close to 40 predictors of AAs that interact with DNA and/or RNA [72-74], over 30 that predict protein-binding AAs

[75, 76]. The sheer number and diversity of these methods make it rather challenging for the end users to select suitable/good predictors and collect their predictions.

The predictive quality of the sequence-based predictors of the AA-level function and structure annotations is evaluated on benchmark datasets. While authors of individual predictors compare their methods to a selected collection of other tools, arguably more reliable information source are community-driven assessments. In the latter case, a large collection of methods competes in a blind prediction task on a common dataset (unknown to the authors of methods) under guidance of an independent group assessors (excluding authors). Examples include the Critical Assessment of Structure Prediction (CASP) [77-79], which evaluates the disorder and contact maps predictions [80, 81], the Critical Assessment of PRotein Interactions (CAPRI) [82-84], Critical Assessment of Intrinsic protein Disorder (CAID) [85], and the discontinued Critical Assessment of Fully Automated Structure Prediction (CAFASP) [86]. The AA-level predictions that do not have community assessments can be reliably compared utilizing large-scale comparative surveys. Recent examples can be found for the prediction of secondary structure [69] (which was discontinued in CASP after 2002), AAs that interact with RNA and DNA [72, 74, 87, 88], protein-binding AAs [75, 89], and disordered protein-binding AAs [76]. The community assessments and the comparative survey give useful guidance for the selection of well-performing predictors.

The collection of predictions could be difficult and time consuming, particularly for less computer savvy users. Users interested in collecting several types of putative annotations have to navigate multiple websites and/or software, correspondingly adjust the format of the input protein sequences, and parse and standardize the diverse formats of outputs that different predictors use. One convenient alternative is to use platforms that provide multiple and diverse predictions. Several platforms that integrate predictions of multiple AA-level descriptors are currently available including PredictProtein [90], PSIPRED workbench [91], MULTICOM [92], Distill [93], and DEPICTER [94]. However, these platforms require a significant amount of runtime to collect results, particularly in scenarios when users require to predict a large number of proteins, and typically focus on a specific annotation type (structural vs. functional) and structural state (disordered vs. structured). Moreover, they are relatively inefficient since the same protein sequence that is being input by different users is typically predicted over and over again.

An ultimate solution to these two prediction problems (selection and collection) are databases that offer convenient access to pre-computed AA-level predictions for a broad collection of predictors. This chapter describes, compares and analyzes these databases in the effort to disseminate and popularize their use.

2 Databases of the AA-level predictions

Three databases of the sequence-based AA-level predictions were released to date: MobiDB [95-98], D^2P^2 [99], and DescribePROT [100]. They provide instantaneous access to results generated by several disorder predictors for large datasets of proteins ranging from 1.35 million proteins in DescribePROT, through 10.43 million proteins in D^2P^2, to 219.74 million proteins in MobiDB. The first two databases focus on annotations associated with intrinsic

disorder while DescribePROT offers a more holistic collection of putative annotations. The disorder is defined by lack of stable structure under physiological conditions [101, 102]. It was bioinformatically shown to be common across all kingdoms of life [103-107] and distributed across cellular compartments [108, 109]. The focus on intrinsic disorder can be explained by its functional importance [110-117], association with human diseases [118] and defining contribution to poorly functionally/structurally characterized dark proteomes [119-121]. The prediction that underly these three databases consists of a numeric propensity (higher value signifies higher likelihood for a given annotation) and a binary value (annotated vs. lacking a given annotation). The binary prediction is typically generated from the propensities, where AAs associated with the propensities higher than a threshold are classified as annotated with a given structural/functional characteristic. We summarize key characteristics of MobiDB, D^2P^2 and DescribePROT in Table 1.

Table 1. Summary of databases of the sequence-based AA-level predictions of protein structure and function.

Database	Refs	Year released	Size [millions of proteins]	Predicted properties: structural (S) and functional (F)	Predictors included	Databases linked
MobiDB version 4.1	[95-98]	2012	219.74	Intrinsic disorder (S) Disordered protein-binding residues (F) Secondary structure (S) Low complexity regions (S) Domains (F)	AlphaFold2 [122] ANCHOR [123] DisEMBL [124] DynaMine [125] ESpritz [126] FeSS [127] Gene3D [128] GlobPlot [129] IUPred2A [38] JRONN [130] MobiDB-lite [131] Pfilt [132] PONDR VSL2B [133, 134] SEG [135]	CoDNaS [136] DIBS [137] DisProt [6] ELM [138] FuzDB [139] IDEAL [140] MFIB [141] PDBe [142] PhasePro [143] UniProt [2]
D^2P^2 version 1.0	[99]	2013	10.43	Intrinsic disorder (S) Disordered protein-binding residues (F) Domains (F)	PONDR VL-XT [144] PONDR VSL2B [133, 134] PrDOS [145] PV2 [146] ESpritz [126] IUPred [40] SUPERFAMILY [147]	IDEAL [140] DisProt [6] PhosphoSitePlus [148]
DescribePROT version 1.4	[100]	2021	1.37	Solvent accessibility (S) Secondary structure (S) Disordered and structured protein-binding (F) Disordered and structured RNA-binding (F) Disordered and structured DNA-binding (F) Intrinsic disorder (S) Disordered linkers (F) Signal peptides (F)	ASAquick [149] DFLpred [150] DRNApred [16] DisoRDPbind [44-46] MoRFChibi [151] PONDR VSL2B [133, 134] PSIPRED [48, 152] SCRIBER [20, 75] SignalP [34, 153]	UniProt [2]

2.1 MobiDB

MobiDB was developed by the Silvio Tosatto's group at the University of Padua. It was first released around 2012 [95], and continues to advance and expand along the years, with version 2 published around 2015 [98], version 3 in 2017 [97], and version 4 in 2020 [96].

Availability: https://mobidb.bio.unipd.it/ [95-98]

Advantages: This is by far the largest database that aims to cover the UniProt-size collection of proteins, which currently totals to 219.7 million. Another key highlight is its linkage to 10 external databases (Table 1) and inclusion of experimental data that was collected from these databases. MobiDB features results generated by 14 predictors, including 8 methods that predict intrinsic disorder. The primary annotation of putative disorder is produced using a meta/consensus method, MobiDB-lite [131]. The meta-predictors input multiple disorder predictions to produce a new disorder prediction that improves over the input predictions. This approach is motivated by empirical works that conclude that well-designed meta-methods in fact produce predictions with favorable accuracy [154, 155].

Disadvantages: MobiDB is almost exclusively focuses on annotations of intrinsic disorder. Moreover, it provides only the binary values for the disorder predictions, lacking the corresponding putative propensities.

2.2 D^2P^2

D^2P^2 was released around 2012 by Julian Gough's team at the University of Bristol. His research group has recently moved to the MRC Laboratory of Molecular Biology at Cambridge and D^2P^2 is no longer supported. The release of this resource was supported by a large international group of researchers including Drs Takeshi Ishida (Tokyo Institute of Technology), Bin Xue and Vladimir Uversky (University of South Florida), Zsuzsanna Dosztanyi (Eotvos Lorand University), Zoran Obradovic (Temple University), Lukasz Kurgan (Virginia Commonwealth University), and A. Keith Dunker (Indiana University).

Availability: https://d2p2.pro/ [99]

Advantages: D^2P^2 offer access to the results produced by a diverse collection of six disorder predictors (Table 1). It also combines these predictions using a 75% consensus approach, i.e., a residue is predicted as disorder if at least 75% of methods predicts it as disordered in binary. The use of this meta/consensus approach is motivated by the past empirical studies [154, 155]. Moreover, D^2P^2 provides arguably the most comprehensive annotations of protein domains and PTMs.

Disadvantages: Similar to MobiDB, D^2P^2 is nearly fully focuses on the intrinsic disorder annotations. Furthermore, this resource was last updated in 2013 and is no longer maintained.

2.3 DescribePROT

DescribePROT was produced by Lukasz Kurgan's lab at the Virginia Commonwealth University and made available to the public in 2020. Similar to D^2P^2, DescribePROT was a collaborative effort that involved a big team of researchers including Drs A. Keith Dunker (Indiana University), Andrzej Kloczkowski (Ohio State University), Jorg Gsponer (University

of British Columbia), Johannes Soding (Max Planck Institute for Biophysical Chemistry), Zoran Obradovic (Temple University), Martin Steinegger (Seoul National University), and Yaoqi Zhou (Shenzhen Bay Laboratory).

Availability: http://biomine.cs.vcu.edu/servers/DESCRIBEPROT/ [100]

Advantages: The strongest point of DescribePROT is the diversity of its predictions that cover several structural and functional characteristics including solvent accessibility, secondary structure, protein-, RNA- and DNA-binding AAs, intrinsic disorder, disordered linkers and signal peptides. Consequently, DescribePROT stores over 7.8 billion AA-level predictions. Moreover, it provides access to position specific scoring matrices (PSSMs) generated from protein sequences using MMSeqs2 [156-158] and the relative entropy-based conservation scores that are produced from PSSMs [159, 160]. Furthermore, this is the only database that combines complementary predictions of DNA, RNA and protein interactions that are trained using structured vs. disordered data [75], which results in a more complete coverage of these interactions.

Disadvantages: The main downside of DescribePROT is a relatively low number of proteins that it covers (1.37 million), which spans over 83 complete proteomes/species. It also suffers insufficient linkage to external resources. However, both of these issues should be resolved in the subsequent releases.

2.4 Example results

Figure 1A shows experimental annotations of structure and function for the SIR3 protein, transcriptional repressor from *Saccharomyces cerevisiae* (UniProt ID: P06701), which we extract from the DisProt database (DisProt ID: DP00533) [6]. SIR3 modulates chromatin structure and correspondingly includes a long intrinsically disordered region (positions 216 to 549) that interacts with proteins and DNA [161].

We compare these annotations against the results that we collect from the D^2P^2 (Figure 1B), MobiDB (Figure 1C) and DescribePROT (Figure 1D) databases. We observe that the location of the predicted disordered AAs in these three databases agrees to a large degree with the experimental data. This suggests that the corresponding disorder predictors produce accurate results, which concurs with recent empirical assessments that similarly conclude that disorder predictions are in general done accurately [85, 162, 163]. We emphasize that these resources provide well-designed and color-coded visualizations of the predictions and annotations, each using its own format. D^2P^2 groups all disorder predictions together and presents an "agreement" line that compares them against experimental annotations, if available (Figure 1B). This is accompanied with the location of identified domains and PTMs. MobiDB similarly clusters several disorder predictions together with the corresponding consensus result (Figure 1C). It also provides annotations of domains and protein interactions at the bottom of the panel. DescribePROT divides the panel into two parts where the top aggregates information at the protein level and the bottom provides complete AA-level results (Figure 1D). The residue-level annotations supplied by DescribePROT include both binary predictions (horizontal bars) and numeric propensities (thin solid lines). We note that MobiDB and DescribePROT provide interactive interfaces where users can select specific

functional/structural characteristics, zoom in and out on selected parts of the sequence, and are shown convenient and informative callouts that display additional details and which appear on the mouse hover.

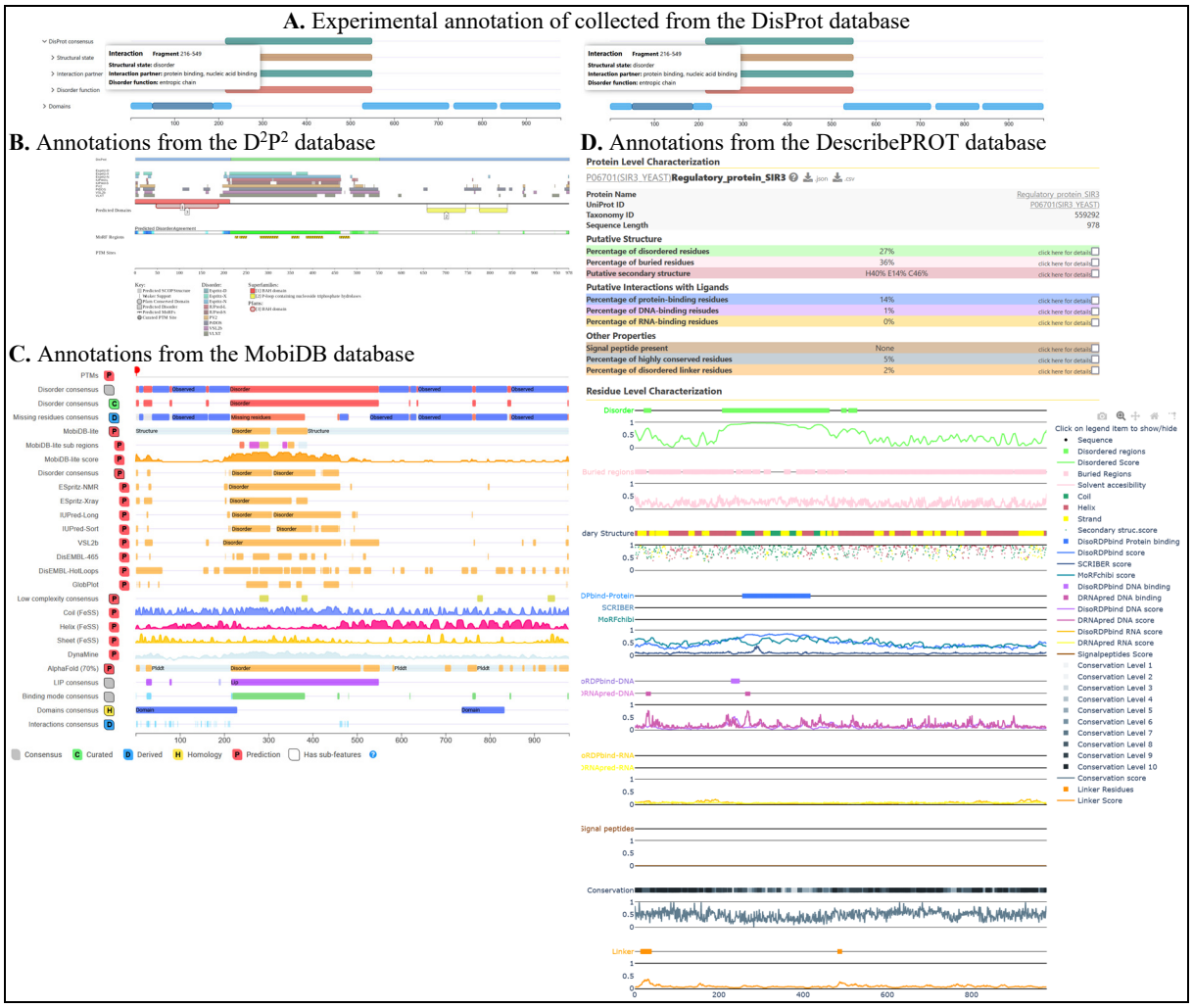

Figure 1. Experimental and predicted disorder annotations for the SIR3 protein (UniProt ID: P06701, DisProt ID: DP00533). Panel A shows the experimental annotations collected from DisProt (https://www.disprot.org/). Panel B shows the results generated by the D^{2}P^2 database (https://d2p2.pro/). Panel C presents the results produced by the MobiDB database (https://mobidb.bio.unipd.it/). Panel D gives the outputs from the DescribePROT database (http://biomine.cs.vcu.edu/servers/DESCRIBEPROT/). The legends included in panels B, C and D explain the encoding of the presented data.

3 Conclusions, impact and limitations

Three large-scale databases that we introduce and discuss in this chapter, MobiDB, D^{2}P^2 and DescribePROT, facilitate easy and free access to large collections of the AA-level annotations of protein structure and function. We demonstrate that they provide complementary scope and services. D^{2}P^2 arguably delivers the most comprehensive set of annotations of protein domains and PTMs. However, this database was last updated in 2013 and is no longer actively

supported. MobiDB focuses primarily on the intrinsic disorder and is by far the largest and most externally connected resource. On the other hand, DescribePROT covers the most diverse collection of the structural and functional features. Thus, we recommend the latter two resources as the most valuable, current and complete solutions to conveniently collect the AA-level annotations.

The data available in these databases is utilized for numerous practical applications. We briefly summarize impact of one of the structural aspects covered by these resources, the intrinsic disorder. Just in 2021, the disorder predictions of the popular IUPred [37-40], which are available via D^2P^2 and MobiDB databases, were used to analyze the SARS-CoV-2 proteins [164-167], link mutations in the intrinsically disordered sequence regions to cancer [168, 169], investigate liquid-liquid phase separation [170-172], localize disorder across compartments of the human cell [108], and to develop a wide range of predictive tools [173-178], among many other applications. Similarly long list of diverse uses can be attributed to the results produced by DisoRDPbind [44-46], which covers putative disordered protein/DNA/RNA binding AAs and which are available via DescribePROT. These predictions were utilized to investigate several viral genomes including SARS-CoV-2 [179], porcine astrovirus type 3 [180], and hepatitis E [181], and to decipher functions of genes from animal pathogens [182]. They were also applied to investigate several specific proteins, such as CS-like zinc finger (FLZ) [183], nonstructural nsP2 protein from Salmonid alphavirus [184], spindle-defective protein 2 (SPD-2) [185], heat shock factor 1 (Hsf1) [186], and Mixed Lineage Leukemia 4 (MLL4) [187], some of which are connected to cancers and neurodegenerative and viral diseases. More broadly, we find that the intrinsic disorder predictions are utilized across many research and development areas, such as drug design [188-192], molecular and systems medicine [193, 194], and structural genomics [124, 195]. These examples and studies clearly demonstrate the significant impact of the use of the putative AA-level annotations, which is directly facilitated by the described here databases.

Lastly, we emphasize that the use of these databases is limited to the proteins that they include. Users who like to collect the AA-level data outside of the protein sets covered in these resources, e.g., for a novel protein sequence, have the option of applying one of the freely available predictive platforms. These platforms include PredictProtein (https://predictprotein.org/) [90], PSIPRED workbench (http://bioinf.cs.ucl.ac.uk/psipred/) [91], MULTICOM (http://sysbio.rnet.missouri.edu/multicom_cluster/) [92], Distill (http://distillf.ucd.ie/distill/) [93], and DEPICTER (http://biomine.cs.vcu.edu/servers/DEPICTER/) [94]. We briefly discuss details of the DisorderEd PredictIon CenTER (DEPICTER) webserver, which is the closest to the scope of the three databases. This webserver conveniently generates the AA-level predictions on the server side covering a broad selection of disorder and disorder function predictions. It produces consensus/meta prediction of disordered AAs using results output by the fast UPred-short [40], IUPred-long [40] and SPOT-Disorder-Single [196] methods. It also predicts the disordered linkers using DFLpred [150], disordered AAs that bind proteins, DNA and/or RNA by combining results of fMoRFpred [197], DisoRDPbind [46] and ANCHOR2 [38] methods, and putative disordered multifunctional (moonlighting) AAs generated by DMRpred [198]. The predictions are visualized and delivered as a parsable text file in the browser window and sent to the user's email, if the email was provided as one of the inputs.

Funding

This work was funded in part by the National Science Foundation (grant 2125218) and the Robert J. Mattauch Endowment funds to L.K.

References

1. Li, W., et al., *RefSeq: expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation.* Nucleic Acids Res, 2021. **49**(D1): p. D1020-D1028.
2. UniProt, C., *UniProt: the universal protein knowledgebase in 2021.* Nucleic Acids Res, 2021. **49**(D1): p. D480-D489.
3. UniProt, C., *UniProt: a hub for protein information.* Nucleic Acids Res, 2015. **43**(Database issue): p. D204-12.
4. wwPDB consortium, *Protein Data Bank: the single global archive for 3D macromolecular structure data.* Nucleic Acids Res, 2019. **47**(D1): p. D520-D528.
5. Yang, J., A. Roy, and Y. Zhang, *BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions.* Nucleic Acids Res, 2013. **41**(Database issue): p. D1096-103.
6. Hatos, A., et al., *DisProt: intrinsic protein disorder annotation in 2020.* Nucleic Acids Res, 2020. **48**(D1): p. D269-D276.
7. Brocchieri, L. and S. Karlin, *Protein length in eukaryotic and prokaryotic proteomes.* Nucleic Acids Research, 2005. **33**(10): p. 3390-3400.
8. Kuznetsov, I.B., et al., *Using evolutionary and structural information to predict DNA-binding sites on DNA-binding proteins.* Proteins-Structure Function and Bioinformatics, 2006. **64**(1): p. 19-27.
9. Hwang, S., Z.K. Gou, and I.B. Kuznetsov, *DP-Bind: a Web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins.* Bioinformatics, 2007. **23**(5): p. 634-636.
10. Ahmad, S. and A. Sarai, *PSSM-based prediction of DNA binding sites in proteins.* BMC Bioinformatics, 2005. **6**: p. 33.
11. Walia, R.R., et al., *RNABindRPlus: a predictor that combines machine learning and sequence homology-based methods to improve the reliability of predicted RNA-binding residues in proteins.* PLoS One, 2014. **9**(5): p. e97725.
12. Terribilini, M., et al., *Prediction of RNA binding sites in proteins from amino acid sequence.* Rna-a Publication of the Rna Society, 2006. **12**(8): p. 1450-1462.
13. Terribilini, M., et al., *RNABindR: a server for analyzing and predicting RNA-binding sites in proteins.* Nucleic Acids Research, 2007. **35**: p. W578-W584.
14. Kumar, M., A.M. Gromiha, and G.P.S. Raghava, *Prediction of RNA binding sites in a protein using SVM and PSSM profile.* Proteins-Structure Function and Bioinformatics, 2008. **71**(1): p. 189-194.
15. Wang, L., et al., *BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features.* BMC Syst Biol, 2010. **4 Suppl 1**: p. S3.
16. Yan, J. and L. Kurgan, *DRNApred, fast sequence-based method that accurately predicts and discriminates DNA- and RNA-binding residues.* Nucleic Acids Res, 2017. **45**(10): p. e84.
17. Su, H., et al., *Improving the prediction of protein-nucleic acids binding residues via multiple sequence profiles and the consensus of complementary methods.* Bioinformatics, 2019. **35**(6): p. 930-936.
18. Porollo, A. and J. Meller, *Prediction-based fingerprints of protein-protein interactions.* Proteins, 2007. **66**(3): p. 630-45.

19. Murakami, Y. and K. Mizuguchi, *Applying the Naive Bayes classifier with kernel density estimation to the prediction of protein-protein interaction sites.* Bioinformatics, 2010. **26**(15): p. 1841-8.
20. Zhang, J. and L. Kurgan, *SCRIBER: accurate and partner type-specific prediction of protein-binding residues from proteins sequences.* Bioinformatics, 2019. **35**(14): p. i343-i353.
21. Katuwawala, A., B. Zhao, and L. Kurgan, *DisoLipPred: Accurate prediction of disordered lipid binding residues in protein sequences with deep recurrent networks and transfer learning.* Bioinformatics, 2021.
22. Song, J., et al., *PROSPERous: high-throughput prediction of substrate cleavage sites for 90 proteases with improved accuracy.* Bioinformatics, 2018. **34**(4): p. 684-687.
23. Li, F., et al., *DeepCleave: a deep learning predictor for caspase and matrix metalloprotease substrates and cleavage sites.* Bioinformatics, 2020. **36**(4): p. 1057-1065.
24. Sankararaman, S. and K. Sjolander, *INTREPID--INformation-theoretic TREe traversal for Protein functional site IDentification.* Bioinformatics, 2008. **24**(21): p. 2445-52.
25. Sankararaman, S., B. Kolaczkowski, and K. Sjolander, *INTREPID: a web server for prediction of functionally important residues by evolutionary analysis.* Nucleic Acids Res, 2009. **37**(Web Server issue): p. W390-5.
26. Song, J.N., et al., *PREvaIL, an integrative approach for inferring catalytic residues using sequence, structural, and network features in a machine-learning framework.* Journal of Theoretical Biology, 2018. **443**: p. 125-137.
27. Zhang, T., et al., *Accurate sequence-based prediction of catalytic residues.* Bioinformatics, 2008. **24**(20): p. 2329-38.
28. Blom, N., et al., *Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence.* Proteomics, 2004. **4**(6): p. 1633-49.
29. Ren, J., et al., *Systematic study of protein sumoylation: Development of a site-specific predictor of SUMOsp 2.0.* Proteomics, 2009. **9**(12): p. 3409-3412.
30. Xue, Y., et al., *SUMOsp: a web server for sumoylation site prediction.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W254-7.
31. Radivojac, P., et al., *Identification, analysis, and prediction of protein ubiquitination sites.* Proteins, 2010. **78**(2): p. 365-80.
32. Nielsen, H., et al., *Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites.* Protein Engineering, 1997. **10**(1): p. 1-6.
33. Bendtsen, J.D., et al., *Improved prediction of signal peptides: SignalP 3.0.* J Mol Biol, 2004. **340**(4): p. 783-95.
34. Almagro Armenteros, J.J., et al., *SignalP 5.0 improves signal peptide predictions using deep neural networks.* Nat Biotechnol, 2019. **37**(4): p. 420-423.
35. Petersen, T.N., et al., *SignalP 4.0: discriminating signal peptides from transmembrane regions.* Nat Methods, 2011. **8**(10): p. 785-6.
36. Emanuelsson, O., H. Nielsen, and G. Von Heijne, *ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites.* Protein Science, 1999. **8**(5): p. 978-984.
37. Erdos, G., M. Pajkos, and Z. Dosztanyi, *IUPred3: prediction of protein disorder enhanced with unambiguous experimental annotation and visualization of evolutionary conservation.* Nucleic Acids Res, 2021. **49**(W1): p. W297-W303.
38. Meszaros, B., G. Erdos, and Z. Dosztanyi, *IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding.* Nucleic Acids Res, 2018. **46**(W1): p. W329-W337.

39. Dosztanyi, Z., *Prediction of protein disorder based on IUPred.* Protein Sci, 2018. **27**(1): p. 331-340.
40. Dosztanyi, Z., et al., *IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content.* Bioinformatics, 2005. **21**(16): p. 3433-4.
41. Jones, D.T. and D. Cozzetto, *DISOPRED3: precise disordered region predictions with annotated protein-binding activity.* Bioinformatics, 2015. **31**(6): p. 857-63.
42. Ward, J.J., et al., *The DISOPRED server for the prediction of protein disorder.* Bioinformatics, 2004. **20**(13): p. 2138-9.
43. Hu, G., et al., *flDPnn: Accurate intrinsic disorder prediction with putative propensities of disorder functions.* Nat Commun, 2021. **12**(1): p. 4438.
44. Oldfield, C.J., Z. Peng, and L. Kurgan, *Disordered RNA-Binding Region Prediction with DisoRDPbind.* Methods Mol Biol, 2020. **2106**: p. 225-239.
45. Peng, Z., et al., *Prediction of Disordered RNA, DNA, and Protein Binding Regions Using DisoRDPbind.* Methods Mol Biol, 2017. **1484**: p. 187-203.
46. Peng, Z. and L. Kurgan, *High-throughput prediction of RNA, DNA and protein binding regions mediated by intrinsic disorder.* Nucleic Acids Res, 2015. **43**(18): p. e121.
47. Zhang, F., et al., *DeepDISOBind: accurate prediction of RNA-, DNA- and protein-binding intrinsically disordered residues with deep multi-task learning.* Brief Bioinform, 2021.
48. McGuffin, L.J., K. Bryson, and D.T. Jones, *The PSIPRED protein structure prediction server.* Bioinformatics, 2000. **16**(4): p. 404-5.
49. Jones, D.T., *Protein secondary structure prediction based on position-specific scoring matrices.* Journal of Molecular Biology, 1999. **292**(2): p. 195-202.
50. Rost, B. and C. Sander, *Prediction of Protein Secondary Structure at Better than 70% Accuracy.* Journal of Molecular Biology, 1993. **232**(2): p. 584-599.
51. Rost, B., C. Sander, and R. Schneider, *Phd - an Automatic Mail Server for Protein Secondary Structure Prediction.* Computer Applications in the Biosciences, 1994. **10**(1): p. 53-60.
52. Cuff, J.A., et al., *JPred: a consensus secondary structure prediction server.* Bioinformatics, 1998. **14**(10): p. 892-3.
53. Cole, C., J.D. Barber, and G.J. Barton, *The Jpred 3 secondary structure prediction server.* Nucleic Acids Res, 2008. **36**(Web Server issue): p. W197-201.
54. Drozdetskiy, A., et al., *JPred4: a protein secondary structure prediction server.* Nucleic Acids Research, 2015. **43**(W1): p. W389-W394.
55. Rost, B. and C. Sander, *Conservation and prediction of solvent accessibility in protein families.* Proteins, 1994. **20**(3): p. 216-26.
56. Pollastri, G., et al., *Prediction of coordination number and relative solvent accessibility in proteins.* Proteins, 2002. **47**(2): p. 142-53.
57. Schlessinger, A., G. Yachdav, and B. Rost, *PROFbval: predict flexible and rigid residues in proteins.* Bioinformatics, 2006. **22**(7): p. 891-3.
58. Schlessinger, A. and B. Rost, *Protein flexibility and rigidity predicted from sequence.* Proteins, 2005. **61**(1): p. 115-26.
59. Chen, K., L.A. Kurgan, and J. Ruan, *Prediction of flexible/rigid regions from protein sequences using k-spaced amino acid pairs.* BMC Struct Biol, 2007. **7**: p. 25.
60. Jones, D.T., et al., *PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments.* Bioinformatics, 2012. **28**(2): p. 184-190.
61. Kamisetty, H., S. Ovchinnikov, and D. Baker, *Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era.* Proceedings of the National Academy of Sciences of the United States of America, 2013. **110**(39): p. 15674-15679.

62. Wang, S., et al., *Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model.* PLoS Comput Biol, 2017. **13**(1): p. e1005324.
63. Cheng, J.L. and P. Baldi, *Improved residue contact prediction using support vector machines and a large feature set.* Bmc Bioinformatics, 2007. **8**.
64. Zhao, B. and L. Kurgan, *Surveying over 100 predictors of intrinsic disorder in proteins.* Expert Rev Proteomics, 2021: p. 1-11.
65. Liu, Y., X. Wang, and B. Liu, *A comprehensive review and comparison of existing computational methods for intrinsically disordered protein and region prediction.* Brief Bioinform, 2019. **20**(1): p. 330-346.
66. Meng, F., V.N. Uversky, and L. Kurgan, *Comprehensive review of methods for prediction of intrinsic disorder and its molecular functions.* Cell Mol Life Sci, 2017. **74**(17): p. 3069-3090.
67. Meng, F., V. Uversky, and L. Kurgan, *Computational Prediction of Intrinsic Disorder in Proteins.* Curr Protoc Protein Sci, 2017. **88**: p. 2 16 1-2 16 14.
68. Kashani-Amin, E., et al., *A systematic review on popularity, application and characteristics of protein secondary structure prediction tools.* Curr Drug Discov Technol, 2018. **16**(2): p. 159-172.
69. Zhang, H., et al., *Critical assessment of high-throughput standalone methods for secondary structure prediction.* Brief Bioinform, 2011. **12**(6): p. 672-88.
70. Oldfield, C.J., K. Chen, and L. Kurgan, *Computational Prediction of Secondary and Supersecondary Structures from Protein Sequences.* Methods Mol Biol, 2019. **1958**: p. 73-100.
71. Meng, F. and L. Kurgan, *Computational Prediction of Protein Secondary Structure from Sequence.* Curr Protoc Protein Sci, 2016. **86**: p. 2 3 1-2 3 10.
72. Yan, J., S. Friedrich, and L. Kurgan, *A comprehensive comparative review of sequence-based predictors of DNA- and RNA-binding residues.* Brief Bioinform, 2016. **17**(1): p. 88-105.
73. Zhang, J., Z. Ma, and L. Kurgan, *Comprehensive review and empirical analysis of hallmarks of DNA-, RNA- and protein-binding residues in protein chains.* Brief Bioinform, 2019. **20**(4): p. 1250-1268.
74. Miao, Z. and E. Westhof, *A Large-Scale Assessment of Nucleic Acids Binding Site Prediction Programs.* PLoS Comput Biol, 2015. **11**(12): p. e1004639.
75. Zhang, J., S. Ghadermarzi, and L. Kurgan, *Prediction of protein-binding residues: dichotomy of sequence-based methods developed using structured complexes versus disordered proteins.* Bioinformatics, 2020. **36**(18): p. 4729-4738.
76. Katuwawala, A., et al., *Computational Prediction of MoRFs, Short Disorder-to-order Transitioning Protein Binding Regions.* Comput Struct Biotechnol J, 2019. **17**: p. 454-462.
77. Alexander, L.T., et al., *Target highlights in CASP14: Analysis of models by structure providers.* Proteins, 2021. **89**(12): p. 1647-1672.
78. Kinch, L.N., et al., *Target classification in the 14th round of the critical assessment of protein structure prediction (CASP14).* Proteins, 2021. **89**(12): p. 1618-1632.
79. Moult, J., *A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction.* Curr Opin Struct Biol, 2005. **15**(3): p. 285-9.
80. Monastyrskyy, B., et al., *Assessment of protein disorder region predictions in CASP10.* Proteins, 2014. **82 Suppl 2**: p. 127-37.
81. Ruiz-Serra, V., et al., *Assessing the accuracy of contact and distance predictions in CASP14.* Proteins, 2021. **89**(12): p. 1888-1900.
82. Lensink, M.F., et al., *Prediction of protein assemblies, the next frontier: The CASP14-CAPRI experiment.* Proteins, 2021. **89**(12): p. 1800-1823.
83. Lensink, M.F., et al., *Modeling protein-protein, protein-peptide, and protein-oligosaccharide complexes: CAPRI 7th edition.* Proteins, 2020. **88**(8): p. 916-938.

84. Dapkunas, J., et al., *Template-based modeling of diverse protein interactions in CAPRI rounds 38-45.* Proteins, 2020. **88**(8): p. 939-947.
85. Necci, M., et al., *Critical assessment of protein intrinsic disorder prediction.* Nat Methods, 2021. **18**(5): p. 472-481.
86. Eyrich, V.A., et al., *CAFASP3 in the spotlight of EVA.* Proteins, 2003. **53 Suppl 6**: p. 548-60.
87. Walia, R.R., et al., *Protein-RNA interface residue prediction using machine learning: an assessment of the state of the art.* BMC Bioinformatics, 2012. **13**: p. 89.
88. Jung, Y., et al., *Partner-specific prediction of RNA-binding residues in proteins: A critical assessment.* Proteins, 2019. **87**(3): p. 198-211.
89. Zhang, J. and L. Kurgan, *Review and comparative assessment of sequence-based predictors of protein-binding residues.* Brief Bioinform, 2018. **19**(5): p. 821-837.
90. Bernhofer, M., et al., *PredictProtein - Predicting Protein Structure and Function for 29 Years.* Nucleic Acids Res, 2021. **49**(W1): p. W535-W540.
91. Buchan, D.W.A. and D.T. Jones, *The PSIPRED Protein Analysis Workbench: 20 years on.* Nucleic Acids Res, 2019. **47**(W1): p. W402-W407.
92. Hou, J., et al., *The MULTICOM protein structure prediction server empowered by deep learning and contact distance prediction*, in *Methods in Mol. Biol.* 2019.
93. Bau, D., et al., *Distill: a suite of web servers for the prediction of one-, two- and three-dimensional structural features of proteins.* BMC Bioinformatics, 2006. **7**: p. 402.
94. Barik, A., et al., *DEPICTER: Intrinsic Disorder and Disorder Function Prediction Server.* J Mol Biol, 2020. **432**(11): p. 3379-3387.
95. Di Domenico, T., et al., *MobiDB: a comprehensive database of intrinsic protein disorder annotations.* Bioinformatics, 2012. **28**(15): p. 2080-2081.
96. Piovesan, D., et al., *MobiDB: intrinsically disordered proteins in 2021.* Nucleic Acids Res, 2021. **49**(D1): p. D361-D367.
97. Piovesan, D., et al., *MobiDB 3.0: more annotations for intrinsic disorder, conformational diversity and interactions in proteins.* Nucleic Acids Res, 2018. **46**(D1): p. D471-D476.
98. Potenza, E., et al., *MobiDB 2.0: an improved database of intrinsically disordered and mobile proteins.* Nucleic Acids Res, 2015. **43**(Database issue): p. D315-20.
99. Oates, M.E., et al., *D(2)P(2): database of disordered protein predictions.* Nucleic Acids Res, 2013. **41**(Database issue): p. D508-16.
100. Zhao, B., et al., *DescribePROT: database of amino acid-level protein structure and function predictions.* Nucleic Acids Res, 2021. **49**(D1): p. D298-D308.
101. Lieutaud, P., et al., *How disordered is my protein and what is its disorder for? A guide through the "dark side" of the protein universe.* Intrinsically Disord Proteins, 2016. **4**(1): p. e1259708.
102. Oldfield, C.J., et al., *Introduction to intrinsically disordered proteins and regions.* Intrinsically Disordered Proteins: Dynamics, Binding, and Function, 2019: p. 1-34.
103. Xue, B., A.K. Dunker, and V.N. Uversky, *Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life.* J Biomol Struct Dyn, 2012. **30**(2): p. 137-49.
104. Peng, Z., et al., *Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life.* Cell Mol Life Sci, 2015. **72**(1): p. 137-51.
105. Wang, C., V.N. Uversky, and L. Kurgan, *Disordered nucleiome: Abundance of intrinsic disorder in the DNA- and RNA-binding proteins in 1121 species from Eukaryota, Bacteria and Archaea.* Proteomics, 2016. **16**(10): p. 1486-98.
106. Peng, Z., M.J. Mizianty, and L. Kurgan, *Genome-scale prediction of proteins with long intrinsically disordered regions.* Proteins, 2014. **82**(1): p. 145-58.

107. Ward, J.J., et al., *Prediction and functional analysis of native disorder in proteins from the three kingdoms of life.* J Mol Biol, 2004. **337**(3): p. 635-45.
108. Zhao, B., et al., *IDPology of the living cell: intrinsic disorder in the subcellular compartments of the human cell.* Cell Mol Life Sci, 2020.
109. Meng, F., et al., *Compartmentalization and Functionality of Nuclear Disorder: Intrinsic Disorder and Protein-Protein Interactions in Intra-Nuclear Compartments.* Int J Mol Sci, 2015. **17**(1).
110. Uversky, V.N., C.J. Oldfield, and A.K. Dunker, *Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling.* Journal of Molecular Recognition, 2005. **18**(5): p. 343-384.
111. Liu, J., et al., *Intrinsic disorder in transcription factors.* Biochemistry, 2006. **45**(22): p. 6873-88.
112. Peng, Z., et al., *A creature with a hundred waggly tails: intrinsically disordered proteins in the ribosome.* Cell Mol Life Sci, 2014. **71**(8): p. 1477-504.
113. Babu, M.M., *The contribution of intrinsically disordered regions to protein function, cellular complexity, and human disease.* Biochem Soc Trans, 2016. **44**(5): p. 1185-1200.
114. Peng, Z.L., et al., *More than just tails: intrinsic disorder in histone proteins.* Molecular Biosystems, 2012. **8**(7): p. 1886-1901.
115. Hu, G., et al., *Functional Analysis of Human Hub Proteins and Their Interactors Involved in the Intrinsic Disorder-Enriched Interactions.* Int J Mol Sci, 2017. **18**(12).
116. Na, I., et al., *Autophagy-related intrinsically disordered proteins in intra-nuclear compartments.* Mol Biosyst, 2016. **12**(9): p. 2798-817.
117. Peng, Z., et al., *Resilience of death: intrinsic disorder in proteins involved in the programmed cell death.* Cell Death Differ, 2013. **20**(9): p. 1257-67.
118. Uversky, V.N., et al., *Unfoldomics of human diseases: linking protein intrinsic disorder with diseases.* BMC Genomics, 2009. **10 Suppl 1**: p. S7.
119. Bhowmick, A., et al., *Finding Our Way in the Dark Proteome.* J Am Chem Soc, 2016. **138**(31): p. 9730-42.
120. Hu, G., et al., *Taxonomic Landscape of the Dark Proteomes: Whole-Proteome Scale Interplay Between Structural Darkness, Intrinsic Disorder, and Crystallization Propensity.* Proteomics, 2018: p. e1800243.
121. Kulkarni, P. and V.N. Uversky, *Intrinsically Disordered Proteins: The Dark Horse of the Dark Proteome.* Proteomics, 2018. **18**(21-22).
122. Jumper, J., et al., *Highly accurate protein structure prediction with AlphaFold.* Nature, 2021. **596**(7873): p. 583-589.
123. Dosztanyi, Z., B. Meszaros, and I. Simon, *ANCHOR: web server for predicting protein binding regions in disordered proteins.* Bioinformatics, 2009. **25**(20): p. 2745-6.
124. Linding, R., et al., *Protein disorder prediction: implications for structural proteomics.* Structure, 2003. **11**(11): p. 1453-9.
125. Cilia, E., et al., *From protein sequence to dynamics and disorder with DynaMine.* Nat Commun, 2013. **4**: p. 2741.
126. Walsh, I., et al., *ESpritz: accurate and fast prediction of protein disorder.* Bioinformatics, 2012. **28**(4): p. 503-9.
127. Piovesan, D., et al., *FELLS: fast estimator of latent local structure.* Bioinformatics, 2017. **33**(12): p. 1889-1891.
128. Lewis, T.E., et al., *Gene3D: Extensive prediction of globular domains in proteins.* Nucleic Acids Res, 2018. **46**(D1): p. D435-D439.
129. Linding, R., et al., *GlobPlot: Exploring protein sequences for globularity and disorder.* Nucleic Acids Res, 2003. **31**(13): p. 3701-8.

130. Yang, Z.R., et al., *RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins.* Bioinformatics, 2005. **21**(16): p. 3369-76.
131. Necci, M., et al., *MobiDB-lite: fast and highly specific consensus prediction of intrinsic disorder in proteins.* Bioinformatics, 2017. **33**(9): p. 1402-1404.
132. Jones, D.T. and M.B. Swindells, *Getting the most from PSI-BLAST.* Trends in Biochemical Sciences, 2002. **27**(3): p. 161-164.
133. Peng, K., et al., *Length-dependent prediction of protein intrinsic disorder.* BMC Bioinformatics, 2006. **7**: p. 208.
134. Obradovic, Z., et al., *Exploiting heterogeneous sequence properties improves prediction of protein disorder.* Proteins, 2005. **61 Suppl 7**: p. 176-82.
135. Wootton, J.C., *Non-globular domains in protein sequences: automated segmentation using complexity measures.* Comput Chem, 1994. **18**(3): p. 269-85.
136. Monzon, A.M., et al., *CoDNaS 2.0: a comprehensive database of protein conformational diversity in the native state.* Database (Oxford), 2016. **2016**.
137. Schad, E., et al., *DIBS: a repository of disordered binding sites mediating interactions with ordered proteins.* Bioinformatics, 2018. **34**(3): p. 535-537.
138. Dinkel, H., et al., *ELM 2016--data update and new functionality of the eukaryotic linear motif resource.* Nucleic Acids Res, 2016. **44**(D1): p. D294-300.
139. Miskei, M., C. Antal, and M. Fuxreiter, *FuzDB: database of fuzzy complexes, a tool to develop stochastic structure-function relationships for protein complexes and higher-order assemblies.* Nucleic Acids Res, 2017. **45**(D1): p. D228-D235.
140. Fukuchi, S., et al., *IDEAL in 2014 illustrates interaction networks composed of intrinsically disordered proteins and their binding partners.* Nucleic Acids Res, 2014. **42**(Database issue): p. D320-5.
141. Ficho, E., et al., *MFIB: a repository of protein complexes with mutual folding induced by binding.* Bioinformatics, 2017. **33**(22): p. 3682-3684.
142. consortium, P.D.-K., *PDBe-KB: collaboratively defining the biological context of structural data.* Nucleic Acids Res, 2022. **50**(D1): p. D534-D542.
143. Meszaros, B., et al., *PhaSePro: the database of proteins driving liquid-liquid phase separation.* Nucleic Acids Res, 2020. **48**(D1): p. D360-D367.
144. Romero, P., et al., *Sequence complexity of disordered protein.* Proteins, 2001. **42**(1): p. 38-48.
145. Ishida, T. and K. Kinoshita, *Prediction of disordered regions in proteins based on the meta approach.* Bioinformatics, 2008. **24**(11): p. 1344-8.
146. Ghalwash, M.F., A.K. Dunker, and Z. Obradovic, *Uncertainty analysis in protein disorder prediction.* Mol Biosyst, 2012. **8**(1): p. 381-91.
147. Gough, J., et al., *Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure.* J Mol Biol, 2001. **313**(4): p. 903-19.
148. Hornbeck, P.V., et al., *PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse.* Nucleic Acids Res, 2012. **40**(Database issue): p. D261-70.
149. Faraggi, E., et al., *Fast and Accurate Accessible Surface Area Prediction Without a Sequence Profile.* Prediction of Protein Secondary Structure, 2017. **1484**: p. 127-136.
150. Meng, F. and L. Kurgan, *DFLpred: High-throughput prediction of disordered flexible linker regions in protein sequences.* Bioinformatics, 2016. **32**(12): p. i341-i350.
151. Malhis, N., M. Jacobson, and J. Gsponer, *MoRFchibi SYSTEM: software tools for the identification of MoRFs in protein sequences.* Nucleic Acids Res, 2016.
152. Buchan, D.W.A. and D.T. Jones, *The PSIPRED Protein Analysis Workbench: 20 years on.* Nucleic Acids Research, 2019. **47**(W1): p. W402-W407.

153. Teufel, F., et al., *SignalP 6.0 predicts all five types of signal peptides using protein language models.* Nat Biotechnol, 2022.
154. Fan, X. and L. Kurgan, *Accurate prediction of disorder in protein chains with a comprehensive and empirically designed consensus.* J Biomol Struct Dyn, 2014. **32**(3): p. 448-64.
155. Peng, Z. and L. Kurgan, *On the complementarity of the consensus-based disorder prediction.* Pac Symp Biocomput, 2012: p. 176-87.
156. Mirdita, M., M. Steinegger, and J. Soding, *MMseqs2 desktop and local web server app for fast, interactive sequence searches.* Bioinformatics, 2019. **35**(16): p. 2856-2858.
157. Steinegger, M. and J. Soding, *MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets.* Nat Biotechnol, 2017. **35**(11): p. 1026-1028.
158. Hu, G. and L. Kurgan, *Sequence Similarity Searching.* Curr Protoc Protein Sci, 2019. **95**(1): p. e71.
159. Wang, K. and R. Samudrala, *Incorporating background frequency improves entropy-based residue conservation measures.* BMC Bioinformatics, 2006. **7**: p. 385.
160. Fischer, J., C. Mayer, and J. Soding, *Prediction of protein functional residues from sequence by probability density estimation.* Bioinformatics, 2008. **24**(5): p. 613-20.
161. McBryant, S.J., C. Krause, and J.C. Hansen, *Domain organization and quaternary structure of the Saccharomyces cerevisiae silent information regulator 3 protein, Sir3p.* Biochemistry, 2006. **45**(51): p. 15941-8.
162. Katuwawala, A. and L. Kurgan, *Comparative Assessment of Intrinsic Disorder Predictions with a Focus on Protein and Nucleic Acid-Binding Proteins.* Biomolecules, 2020. **10**(12).
163. Katuwawala, A., C.J. Oldfield, and L. Kurgan, *Accuracy of protein-level disorder predictions.* Brief Bioinform, 2020. **21**(5): p. 1509-1522.
164. Schiavina, M., et al., *The highly flexible disordered regions of the SARS-CoV-2 nucleocapsid N protein within the 1-248 residue construct: sequence-specific resonance assignments through NMR.* Biomol NMR Assign, 2021. **15**(1): p. 219-227.
165. Zhao, M., et al., *GCG inhibits SARS-CoV-2 replication by disrupting the liquid phase condensation of its nucleocapsid protein.* Nat Commun, 2021. **12**(1): p. 2114.
166. Cubuk, J., et al., *The SARS-CoV-2 nucleocapsid protein is dynamic, disordered, and phase separates with RNA.* Nature Communications, 2021. **12**(1).
167. Akbayrak, I.Y., et al., *Structures of MERS-CoV macro domain in aqueous solution with dynamics: Impacts of parallel tempering simulation techniques and CHARMM36m and AMBER99SB force field parameters.* Proteins, 2021. **89**(10): p. 1289-1299.
168. Meszaros, B., et al., *Mutations of Intrinsically Disordered Protein Regions Can Drive Cancer but Lack Therapeutic Strategies.* Biomolecules, 2021. **11**(3).
169. Wong, E.T.C., et al., *Protein-Protein Interactions Mediated by Intrinsically Disordered Protein Regions Are Enriched in Missense Mutations.* Biomolecules, 2020. **10**(8).
170. Agarwal, A., et al., *An intrinsically disordered pathological prion variant Y145Stop converts into self-seeding amyloids via liquid-liquid phase separation.* Proc Natl Acad Sci U S A, 2021. **118**(45).
171. Li, J., et al., *Protein phase separation and its role in chromatin organization and diseases.* Biomed Pharmacother, 2021. **138**: p. 111520.
172. Agarwal, A. and S. Mukhopadhyay, *Prion Protein Biology Through the Lens of Liquid-Liquid Phase Separation.* Journal of Molecular Biology, 2022. **434**(1): p. 167368.
173. Katuwawala, A., et al., *QUARTERplus: Accurate disorder predictions integrated with interpretable residue-level quality assessment scores.* Comput Struct Biotechnol J, 2021. **19**: p. 2597-2606.

174. Katuwawala, A., C.J. Oldfield, and L. Kurgan, *DISOselect: Disorder predictor selection at the protein level.* Protein Sci, 2020. **29**(1): p. 184-200.
175. Ghadermarzi, S., et al., *XRRpred: Accurate Predictor of Crystal Structure Quality from Protein Sequence.* Bioinformatics, 2021.
176. Zhang, J., et al., *DNAgenie: accurate prediction of DNA-type-specific binding residues in protein sequences.* Brief Bioinform, 2021. **22**(6).
177. van Mierlo, G., et al., *Predicting protein condensate formation using machine learning.* Cell Rep, 2021. **34**(5): p. 108705.
178. Lei, Y., et al., *A deep-learning framework for multi-level peptide-protein interaction prediction.* Nat Commun, 2021. **12**(1): p. 5465.
179. Giri, R., et al., *Understanding COVID-19 via comparative analysis of dark proteomes of SARS-CoV-2, human SARS and bat SARS-like coronaviruses.* Cell Mol Life Sci, 2020.
180. Matias Ferreyra, F., et al., *Comparative Analysis of Novel Strains of Porcine Astrovirus Type 3 in the USA.* Viruses, 2021. **13**(9).
181. Zoya, S., et al., *Shedding light on the dark proteome of Hepatitis E Virus.* Network Biology, 2021. **11**(4): p. 295-314.
182. Oliva Chavez, A.S., et al., *Mutational analysis of gene function in the Anaplasmataceae: Challenges and perspectives.* Ticks Tick Borne Dis, 2019. **10**(2): p. 482-494.
183. Jamsheer, K.M., et al., *The FCS-like zinc finger scaffold of the kinase SnRK1 is formed by the coordinated actions of the FLZ domain and intrinsically disordered regions.* J Biol Chem, 2018. **293**(34): p. 13134-13150.
184. Jami, R., et al., *The C-Terminal Domain of Salmonid Alphavirus Nonstructural Protein 2 (nsP2) Is Essential and Sufficient To Block RIG-I Pathway Induction and Interferon-Mediated Antiviral Response.* J Virol, 2021. **95**(23): p. e0115521.
185. Murph, M., S. Singh, and M. Schvarzstein, *The Centrosomal Swiss Army Knife: A combined in silico and in vivo approach to the structure-function annotation of SPD-2 provides mechanistic insight into its functional diversity.* bioRxiv, 2021: p. 2021.04.22.441031.
186. Pujols, J., et al., *The Disordered C-Terminus of Yeast Hsf1 Contains a Cryptic Low-Complexity Amyloidogenic Region.* Int J Mol Sci, 2018. **19**(5).
187. Szabo, B., et al., *Disordered Regions of Mixed Lineage Leukemia 4 (MLL4) Protein Are Capable of RNA Binding.* Int J Mol Sci, 2018. **19**(11).
188. Hu, G., et al., *Untapped Potential of Disordered Proteins in Current Druggable Human Proteome.* Curr Drug Targets, 2016. **17**(10): p. 1198-205.
189. Hosoya, Y. and J. Ohkanda, *Intrinsically Disordered Proteins as Regulators of Transient Biological Processes and as Untapped Drug Targets.* Molecules, 2021. **26**(8).
190. Biesaga, M., M. Frigole-Vivas, and X. Salvatella, *Intrinsically disordered proteins and biomolecular condensates as drug targets.* Curr Opin Chem Biol, 2021. **62**: p. 90-100.
191. Ambadipudi, S. and M. Zweckstetter, *Targeting intrinsically disordered proteins in rational drug discovery.* Expert Opin Drug Discov, 2016. **11**(1): p. 65-77.
192. Ghadermarzi, S., et al., *Sequence-Derived Markers of Drug Targets and Potentially Druggable Human Proteins.* Front Genet, 2019. **10**: p. 1075.
193. Deng, X., et al., *An Overview of Practical Applications of Protein Disorder Prediction and Drive for Faster, More Accurate Predictions.* Int J Mol Sci, 2015. **16**(7): p. 15384-404.
194. Kurgan, L., M. Li, and Y. Li, *The Methods and Tools for Intrinsic Disorder Prediction and their Application to Systems Medicine*, in *Systems Medicine*, O. Wolkenhauer, Editor. 2021, Academic Press: Oxford. p. 159-169.
195. Oldfield, C.J., et al., *Utilization of protein intrinsic disorder knowledge in structural proteomics.* Biochim Biophys Acta, 2013. **1834**(2): p. 487-98.

196. Hanson, J., K. Paliwal, and Y. Zhou, *Accurate Single-Sequence Prediction of Protein Intrinsic Disorder by an Ensemble of Deep Recurrent and Convolutional Architectures.* J Chem Inf Model, 2018. **58**(11): p. 2369-2376.
197. Yan, J., et al., *Molecular recognition features (MoRFs) in three domains of life.* Mol Biosyst, 2016. **12**(3): p. 697-710.
198. Meng, F. and L. Kurgan, *High-throughput prediction of disordered moonlighting regions in protein sequences.* Proteins, 2018. **86**(10): p. 1097-1110.