Nucleic Acids Research, 2023, 1–8
https://doi.org/10.1093/nar/gkad985
Database issue

OXFORD

DescribePROT in 2023: more, higher-quality and experimental annotations and improved data download options

Sushmita Basu[1,†], Bi Zhao[2,†], Bálint Biró[1,3], Eshel Faraggi[4], Jörg Gsponer[5], Gang Hu[6], Andrzej Kloczkowski[7], Nawar Malhis[5], Milot Mirdita[8], Johannes Söding[9], Martin Steinegger[8,10,11], Duolin Wang[12], Kui Wang[6], Dong Xu[12], Jian Zhang[13] and Lukasz Kurgan[1,*]

[1]Department of Computer Science, Virginia Commonwealth University, Richmond, VA, USA
[2]Genomics Program, College of Public Health, University of South Florida, Tampa, FL, USA
[3]Department of Animal Biotechnology, Hungarian University of Agriculture and Life Sciences, Gödöllő, Hungary
[4]Physics Department, Indiana University, Indianapolis, IN, USA
[5]Michael Smith Laboratories, University of British Columbia, Vancouver, British Columbia, Canada
[6]School of Statistics and Data Science, LPMC and KLMDASR, Nankai University, Tianjin, P.R. China
[7]The Steve and Cindy Rasmussen Institute for Genomic Medicine, Nationwide Children's Hospital, Columbus, USA
[8]School of Biological Sciences, Seoul National University, Seoul, Republic of Korea
[9]Quantitative and Computational Biology, Max Planck Institute for Multidisciplinary Sciences, Göttingen, Germany
[10]Institute of Molecular Biology & Genetics, Seoul National University, Seoul, Republic of Korea
[11]Artificial Intelligence Institute, Seoul National University, Seoul, South Korea
[12]Department of Electrical Engineer and Computer Science, Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, USA
[13]School of Computer and Information Technology, Xinyang Normal University, Xinyang, P.R. China

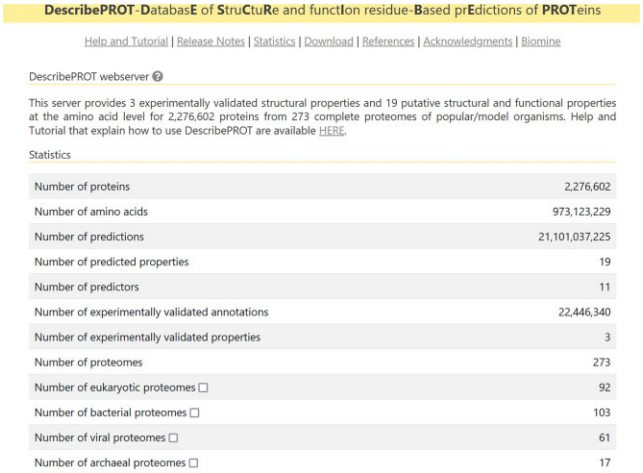
[*]To whom correspondence should be addressed. Tel: +1 804 827 3986; Email: lkurgan@vcu.edu.
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Abstract

The DescribePROT database of amino acid-level descriptors of protein structures and functions was substantially expanded since its release in 2020. This expansion includes substantial increase in the size, scope, and quality of the underlying data, the addition of experimental structural information, the inclusion of new data download options, and an upgraded graphical interface. DescribePROT currently covers 19 structural and functional descriptors for proteins in 273 reference proteomes generated by 11 accurate and complementary predictive tools. Users can search our resource in multiple ways, interact with the data using the graphical interface, and download data at various scales including individual proteins, entire proteomes, and whole database. The annotations in DescribePROT are useful for a broad spectrum of studies that include investigations of protein structure and function, development and validation of predictive tools, and to support efforts in understanding molecular underpinnings of diseases and development of therapeutics. DescribePROT can be freely accessed at http://biomine.cs.vcu.edu/servers/DESCRIBEPROT/.

Received: September 14, 2023. **Revised:** October 12, 2023. **Editorial Decision:** October 13, 2023. **Accepted:** October 16, 2023

Graphical abstract



Introduction

With millions of protein sequences that have been collected (1), we are confronted with the great challenge of characterizing them functionally and structurally. These annotations are done at three levels: the atomic, amino acid (AA) and whole protein level. The notable atomic-level databases include Protein Data Bank (PDB) (2,3), the primary depository of experimentally solved structures, and AlphaFold DB (4), which houses putative structures predicted with AlphaFold2 (5). On the other end of the spectrum, the primary protein-level database, UniProtKB, consists of two parts: 570 thousand proteins that have undergone manual review (Swiss-Prot) and roughly 248 million computationally annotated proteins (TrEMBL) (6,7). The AA-level annotations, also named as one-dimensional descriptors, bridge the gap between the atomic and protein-level data by describing structural and functional features of AAs that compose protein chains (8,9). The structural descriptors include solvent accessibility, secondary structure, linkers, intrinsic disorder, and flexibility. Commonly used functional descriptors cover annotations of protein domains, catalytic residues, and residues that interact with specific types of partners, such as proteins, peptides, nucleic acids, and lipids. While these AA-level annotations can be computed from the atomic structure files and collected from Swiss-Prot/TrEMBL records, they cover a relatively small subset of proteins in the case of PDB and a small subset of AAs in the Swiss-Prot/TrEMBL annotated sequences.

The AA-level descriptors can be predicted from the sequences using hundreds of available computational methods (10–20). However, selecting a sufficiently fast and accurate collection of relevant methods is challenging, and using them in tandem is rather difficult (i.e. different websites and software must be used, and their results have to be reformatted) and wasteful (i.e. different users would make the same predictions when studying the same proteins). Some of these issues can be mitigated with the help of predictive platforms that provide integrated access to multiple predictors, such as PSIPRED workbench (21,22), MULTICOM toolbox (23,24), DEPICTER web server (25,26) and LambdaPP service (27,28). An alternative solution is offered by three databases of the AA-level annotations: D^2P^2 (29), MobiDB (30–32) and DescribePROT (33). These resources provide quick and convenient access to pre-computed results generated by multiple predictors and some experimental annotations. Both options are subject to certain trade-offs. The databases are limited to the proteins that they cover while the predictive platforms can be used for virtually any protein sequence. On the other hand, these platforms require a relatively considerable amount of runtime to complete predictions, especially when processing large collection of proteins, and make repeated predictions, while databases provide fast access to predictions for even larger protein sets. Among the existing databases, MobiDB and D^2P^2 have a nearly singular focus on the intrinsic disorder, and the latter was last updated in 2012 and is no longer maintained. At the point of its release in May 2020, DescribePROT provided access to the predictions of 10 descriptors generated by 10 predictors for a collection of 1.37 million proteins from 83 complete proteomes of popular organisms. Data from DescribePROT found applications in a broad range of contexts, including the development of new predictors of protein function (34,35), investigations of protein functions (36–38), drug design efforts (39), and studies of molecular underpinnings of human diseases (40,41). Since the initial release, our resource underwent eight major updates that collectively expanded its coverage to 19 putative descriptors for 2.28 million proteins from 273 proteomes, incorporated experimental data, provided new options to download data, and introduced an improved graphical interface.

Database description

DescribePROT is freely available at http://biomine.cs.vcu.edu/servers/DESCRIBEPROT/. We provide three convenient ways to search our database: by UniProt accession number, UniProt entry name, and FASTA-formatted sequence. The latter search generates a collection of proteins sorted by similarity to the input chain, which we produce using BLAST (55,56). They are, by default, sorted using the E-values, but the protein list can be resorted using other measures, such as identity and coverage. Proteins on the sorted list are identified by their names, taxonomy IDs, and accession numbers that link to their UniProt records, providing helpful context for selecting the most relevant protein. In case the user-provided accession number or entry name cannot be found, we redirect

Table 1. Comparison of the scope of the initial release 1.0 of DescribePROT and the most recent version 2.0

	DescribePROT 1.0	DescribePROT 2.0
Number of proteomes	83	273
Number of proteins	1 365 946	2 276 602
Number of predictions	7 793 698 419	21 101 037 225
Number and list of predicted descriptors	10 Structural (4): solvent accessibility, secondary structure, intrinsic disorder, and disordered linkers Functional (4): protein-binding, MoRFs, RNA-binding, and DNA-binding Sequence (2): sequence conservation and signal peptides	19 Structural (4): solvent accessibility, secondary structure, intrinsic disorder, and disordered linkers Functional (13): protein-binding, MoRFs, RNA-binding, DNA-binding; phosphorylation, glycosylation, ubiquitination, SUMOylation, acetylation, methylation, pyrrolidone carboxylic acid, palmitoylation and hydroxylation Sequence (2): Sequence conservation and signal peptides
Number and list of predictive methods used	10 ASAquick (42); DFLpred (43); DisoRDPbind (44); DRNApred (45); MMseqs2 (46,47); MoRFchiBi (48); PSIPRED (21,22); SCRIBER (49); SignalP (50); VSL2B (51)	11 ASAquick (42); DFLpred (43); DisoRDPbind (44); DRNApred (45); flDPnn (52); MMseqs2 (46,47); MoRFchiBi (48); MusiteDeep (53,54); PSIPRED (21,22); SCRIBER (49); SignalP (50)
Number of experimental annotations	0	2 244 6340
Number and list of experimental descriptors	0	3 Structural (3): solvent accessibility, secondary structure, and intrinsic disorder

Table 2. Summary and taxonomic classification of protein data and predictions included in DescribePROT

Taxonomic classification		No. of proteomes	No. of sequences	No. of AAs	No. of predictions
Eukaryotes	Animalia	41	893 034	434 456 143	9 349 965 970
	Plantae	17	606 916	232 930 949	5 046 831 941
	Fungi	17	124 010	60 280 917	1 322 327 409
	Protista	17	222 477	109 609 581	2 399 560 661
Bacteria		103	384 907	122 975 734	2 703 915 554
Archaea		17	39 207	11 140 376	240 727 024
Viruses		61	6051	1 729 529	37 708 666
Total		**273**	**2 276 602**	**973 123 229**	**21 101 037 225**

users to identify the most similar protein using the sequence search. DescribePROT also provides direct landing pages for specific proteins that are identified with the UniProt accession numbers, e.g. results for P04637 (p53 protein) can be fetched with the following direct link: http://biomine.cs.vcu.edu/servers/DESCRIBEPROT/result_v2.php?uniprot=P04637.

Table 1 summarizes major changes since the initial release of DescribePROT, which we describe in the following subsections.

Increased coverage of proteomes

We expanded the original list of 83 proteomes to 273 proteomes (230% increase), which we identified by using the '(proteome_type:1) AND (cpd:4)' search in the UniProt database. This search outputs reference proteomes with high levels of completeness of their protein sets that provide broad coverage of the Tree of Life and include frequently sought-after model organisms and other proteomes of interest for biomedical and biotechnological research. We increased the original list of organisms from 56 to 92 eukaryotes, 8 to 103 bacteria, 6 to 17 archaea, and 13 to 61 viruses, with the underlying aim to provide a more taxonomically balanced and inclusive coverage that encompasses all major model organisms (e.g. human, mouse, rat, zebrafish, macaque, fruit fly, yeast, *Caenorhabditis elegans*, *Arabidopsis thaliana*, and *Escherichia coli*) and viruses (e.g. SARS-CoV-2, HIV, papillomavirus, influenza, Ebola, mumps, and herpes). Correspondingly, this resulted in a significant growth in the number of included proteins from 1.36 million to 2.28 million, and in the total number of AA-level predictions from 7.7 billion to 21.1 billion. Table 2 and Figure 1 provide a more detailed taxonomic breakdown of these data. Figure 1A shows that DescribePROT includes 16.9% bacterial proteins, 1.7% archaeal proteins, 0.3% viral proteins and 81.1% eukaryotic proteins. For comparison, the fraction of eukaryotic proteins in the original version of DescribePROT was 96.6%. Moreover, Figure 1B reveals that the distribution of the proteomes is relatively balanced with 33.7% eukaryotes, 37.7% bacteria, 22.3% viruses and 6.3% archaea. The disproportionally larger number of eukaryotic proteins is due to the large sizes of these proteomes.

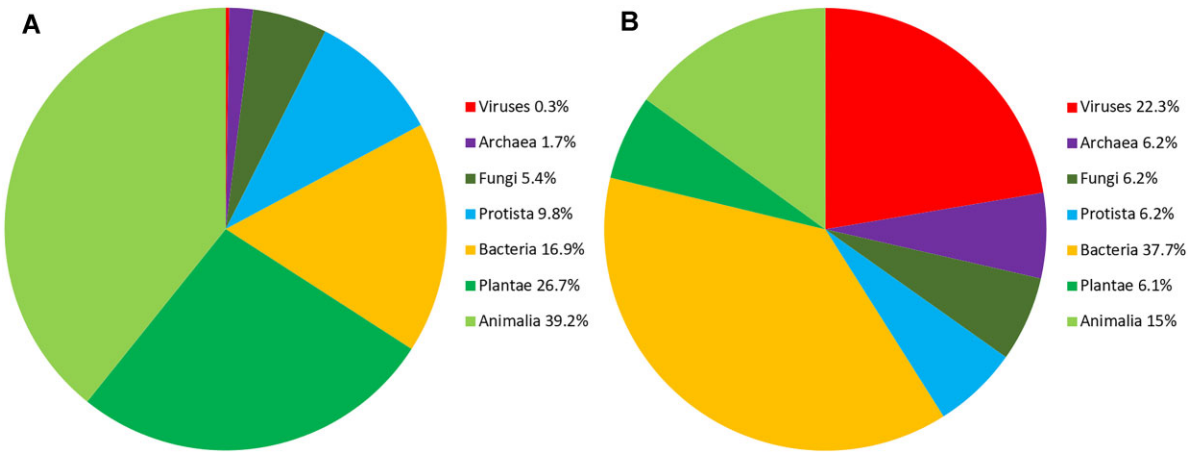


Figure 1. Taxonomic distribution of the proteins (panel **A**) and proteomes (panel **B**) in DescribePROT.

Improved coverage and quality of the AA-level descriptors

We carefully select tools that we use to generate putative annotations in DescribePROT. They must satisfy four key characteristics: (i) relatively short runtime (up to on average 100 s per protein), which is necessary given the large number of proteins; (ii) availability of a working standalone code that we can run locally; (iii) accurate predictions and (iv) complementary coverage of a comprehensive set of AA-level descriptors. Details and justification for the selection of the initial group of the 10 predictors (Table 1) are included in the original DescribePROT article (33). The new DescribePROT implements several key improvements by introducing a new predictor that generates putative annotations of several types of posttranslational modifications (PTMs), MusiteDeep (53,54), replacing the VLS2B disorder predictor (51) with a more modern and accurate flDPnn method (52), and improving quality of the sequence conservation scores generated with MMseqs2. Altogether, the current version of DescribePROT covers 19 diverse descriptors, which cover 4 structural, 13 functional and 2 sequence characteristics (Table 1). The 'Methods' page that is linked at the top of the DescribePROT's main web server page briefly summarizes the 11 included tools and discusses quality of their results. We provide further details for the new and improved tools in the following three paragraphs.

MusiteDeep is a fast, accurate and popular predictor of PTMs that satisfies the above selection criteria (53,54). It applies modular design where separate hybrid deep networks composed of convolutional and fully connected feed-forward layers are used to predict different PTM types. These networks were trained using transfer learning and bootstrapping techniques to maximize predictive performance, achieving an average area under the ROC curve (AUC) over different PTM types of 0.931 (53). As a point of reference, another popular tool, ModPred (57), secured an average AUC of 0.754 in the same test (53). MusiteDeep calculates a numeric score representing the likelihood of a specific PTM type and assigns binary labels (PTM site) to the corresponding amino acid types in the protein sequence. We use MusiteDeep to predict nine major types of PTMs: phosphorylation, glycosylation, ubiquitination, SUMOylation, acetylation, methylation, hydroxylation, pyrrolidone carboxylic acid and palmitoylation.

Replacement of VSL2B with flDPnn is motivated by the results of a large community-run CAID (Critical Assessment of protein Intrinsic Disorder prediction) experiment (58). The flDPnn method was found to be the fastest among the most accurate disorder predictors in CAID (58,59). It secures per-protein runtime of 25 seconds and outperforms VSL2B by a wide margin, with AUCs of 0.814 versus 0.732 on the DisProt dataset in CAID (58). This large margin of improvement was confirmed in a subsequent study (60). flDPnn relies on a deep feed-forward neural network that uses custom-designed inputs derived from the sequence that include sequence conservation, putative secondary structure, and initial predictions of disorder, disordered linkers and disordered binding. It produces a numeric propensity for intrinsic disorder and a binary label (disordered vs. structured) for each AA in the protein sequence.

We also improved quality of the sequence conservation values. As previously, we applied MMseqs2 (46,47) to generate position specific scoring matrices (PSSMs) using the relative entropy-based approach (61), with the background amino acid frequencies derived from BLOSUM62 (62). However, we now calculate the PSSMs using the substantially larger UniRef90 (63) dataset collected from the 2022_03 release of UniProt (151 million proteins), as compared to the reference proteomes dataset from the 2019_08 release that we used previously (56 million proteins).

Moreover, we introduce experimental annotations of the secondary structure, solvent accessibility, and intrinsic disorder that we collected and aggregated from the source data in the PDB (2,3) (for structured proteins) and DisProt (64) (for intrinsically disordered proteins) databases. We rely on the DSSP software (65) to compute secondary structure and solvent accessibility from the PDB structures, normalize the DSSP-derived absolute solvent accessibility using the residue-specific factors from ref. (66) to obtain relative solvent accessibility, and apply in-house scripts to extract annotations of disorder from the DisProt files. Moreover, we follow the protocol from ref. (67) to map data collected from the PDB chains into the corresponding UniProt sequences. This protocol maps multiple structures that correspond to the same protein to increase coverage. Altogether, we added 22446340 experimental annotations for these three descriptors.

Figure 2. An example graphical page with data for the human p53 protein (UniProt ID: P04637). Red callouts identify key features of this page.

Improved accessibility

The data from DescribePROT are available to the users in multiple convenient and complementary formats and levels of aggregation, including individual protein, whole proteome, and the entire database. We provide the raw source data (all predictions and experimental annotations) in the JSON format for each of the 273 proteomes and the entire database. We also introduce new protein-level aggregates of the underlying 19 AA-level descriptors for each proteome in an easy-to-parse CSV format. These aggregates can be used to quickly assess and compare key structural and functional characteristics of individual proteins. They include helix, strand and coil content (% of helix, strand and coil residues); content of buried residues (% of amino acids that are buried in the structure); intrinsic disorder content; content of residues in linker regions; content of residues that bind RNA, DNA, proteins, and peptides; content of residues that are conserved and that have low conservation scores; content of the nine types of PTM sites; and presence of a signal peptide. The downloadable files are available at http://biomine.cs.vcu.edu/servers/DESCRIBEPROT/download.html.

We provide access to the data for individual proteins via an interactive graphical page (Figure 2). The top of the page includes the accession number (linked to the corresponding UniProt record) and links to the CSV- and JSON-formatted files with the raw data. This is followed by protein information panel that provides protein name, taxonomy identifier and length of the sequence. Next, we include color-coded descriptors that aggregate structural and function information at the protein level, providing a quick overview of key characteristics of the selected protein. The bottom portion of the page is an interactive graphical panel that gives residue-level descriptors and applies the same color schema as the upper panel. We summarize each descriptor with two plots that include horizontal bars that identify sequence regions and immediately below a line plot that visualizes numeric propensities/scores for each amino acid in the sequence. The regions are composed of residues with high propensity scores. We redesigned the original graphical panel to provide a better layout, callouts that appear on mouse hover showing useful residue-level and sequence region-level summaries, and new interactive features. Users can hide/show individual plots (by clicking on their name in the menu on the right side), zoom in on specific sequence regions, produce a spike line to identify position and underlying values for individual amino acids, and pan (move horizontally) the panel. The graphical view can be saved as an

image in the PNG format by clicking the little camera icon in the menu on the right side.

We illustrate some of the above-mentioned features using results for the human cellular tumor antigen p53 (UniProt ID: P04637) in Figure 2. This protein is involved in several key cellular processes, such as apoptosis and DNA repair (68), and has multiple intrinsically disordered regions that interact with a large number of proteins (69–73) and DNA (74,75) partners. The top panel with protein-level summaries shows that the native/experimental disorder content is 38% while the prediction suggests that 47% of residues are disordered (green annotations in Figure 2). The native disordered regions are shown using the dark green horizontal bars at the top of the graphical panel in the middle of Figure 2. The callout reveals that the disordered region at the N-terminus stretches between positions 1 and 93. The light green plots immediately below are the disorder predictions generated by the flDPnn method, which are in good agreement with the experimentally identified regions. The blue- and purple-colored annotations focus on interactions with proteins and DNA, respectively. The top panel suggests that 20% of residues are involved in the interactions with proteins and 4% with DNA, which is in line with the experimental data (71).

Discussion

DescribePROT improves on its previous version across all key aspects by enlarging size and scope, improving the quality of underlying data, adding experimental information, providing new data download options, and revamping and improving the graphical interface. Our resource currently covers 19 structural and functional characteristics for proteins from 273 reference proteomes, focusing on model organisms and species of interest for biomedical research that provide taxonomically balanced coverage of the Tree of Life. We also offer multiple ways to conveniently search, download and interact with the data.

We aim to update DescribePROT quarterly. Future changes will primarily concentrate on further expanding the size, coverage and functionality of our resource. We plan to continually grow the number of included proteomes and introduce a broader selection of experimental annotations. The latter will draw on collection and processing of data from a number of relevant databases that include PhosphoSitePlus (76) for PTMs, InterPro (77) for protein domains and BioLip (78) for annotations of protein-ligand interactions. We are also actively working on adding structural data derived from the AlphaFold2 DB (4). We plan to use the putative tertiary structures produced by AlphaFold2 to extract the secondary structure and solvent accessibility descriptors, eventually replacing the currently used programs. We note that dedicated disorder predictors were shown to provide more accurate results than AlphaFold2 for the disorder predictions (79,80) while other functional and structural characteristics cannot be extracted from AlphaFold2 DB. A current impediment that prevents us from using the AlphaFold2 DB data is that they do not cover viral proteomes. Moreover, we are in the process of implementing Application Programming Interface (API) access to facilitate programmable interactions with our resource. We are applying for networking permissions to setup this access and we anticipate that this work could be completed by early 2024.

Data availability

The database and all corresponding data are freely available at http://biomine.cs.vcu.edu/servers/DESCRIBEPROT/.

Acknowledgements

We gratefully acknowledge the contributions of the authors of the predictive tools that were used to develop this resource, which were developed by the labs of Drs Jörg Gsponer, David T Jones, Andrzej Kloczkowski, Lukasz Kurgan, Henrik Nielsen, Johannes Söding and Dong Xu. We also thank users of our resource for their constructive feedback that motivated some of the improvements to our resource.

Funding

National Science Foundation [2146027, 2125218 to L.K., in part]; Robert J. Mattauch Endowment funds (to L.K.); National Institutes of Health [R35-GM126985 to D.X.]. Funding for open access charge: National Science Foundation.

Conflict of interest statement

None declared.

References

1. Li,W., O'Neill,K.R., Haft,D.H., DiCuccio,M., Chetvernin,V., Badretdin,A., Coulouris,G., Chitsaz,F., Derbyshire,M.K., Durkin,A.S., *et al.* (2021) RefSeq: expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation. *Nucleic Acids Res.*, **49**, D1020–D1028.
2. Burley,S.K., Bhikadiya,C., Bi,C., Bittrich,S., Chen,L., Crichlow,G.V., Christie,C.H., Dalenberg,K., Di Costanzo,L., Duarte,J.M., *et al.* (2021) RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Res.*, **49**, D437–D451.
3. Burley,S.K., Bhikadiya,C., Bi,C., Bittrich,S., Chao,H., Chen,L., Craig,P.A., Crichlow,G.V., Dalenberg,K., Duarte,J.M., *et al.* (2023) RCSB Protein Data Bank (RCSB.org): delivery of experimentally-determined PDB structures alongside one million computed structure models of proteins from artificial intelligence/machine learning. *Nucleic Acids Res.*, **51**, D488–D508.
4. Varadi,M., Anyango,S., Deshpande,M., Nair,S., Natassia,C., Yordanova,G., Yuan,D., Stroe,O., Wood,G., Laydon,A., *et al.* (2022) AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.*, **50**, D439–D444.
5. Jumper,J., Evans,R., Pritzel,A., Green,T., Figurnov,M., Ronneberger,O., Tunyasuvunakool,K., Bates,R., Zidek,A., Potapenko,A., *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**, 583–589.
6. Boutet,E., Lieberherr,D., Tognolli,M., Schneider,M., Bansal,P., Bridge,A.J., Poux,S., Bougueleret,L. and Xenarios,I. (2016) UniProtKB/Swiss-Prot, the manually annotated section of the UniProt KnowledgeBase: how to use the entry view. *Methods Mol. Biol.*, **1374**, 23–54.
7. The UniProt Consortium (2023) UniProt: the Universal Protein knowledgebase in 2023. *Nucleic Acids Res.*, **51**, D523–D531.
8. Kurgan,L. and Disfani,F.M. (2011) Structural protein descriptors in 1-dimension and their sequence-based predictions. *Curr. Protein Pept. Sci.*, **12**, 470–489.

9. Rost,B. (2003) Prediction in 1D: secondary structure, membrane helices, and accessibility. *Methods Biochem. Anal.*, **44**, 559–587.
10. Zhao,B. and Kurgan,L. (2021) Surveying over 100 predictors of intrinsic disorder in proteins. *Expert Rev. Proteomics*, **18**, 1019–1029.
11. Liu,Y., Wang,X. and Liu,B. (2019) A comprehensive review and comparison of existing computational methods for intrinsically disordered protein and region prediction. *Brief. Bioinform.*, **20**, 330–346.
12. Basu,S., Kihara,D. and Kurgan,L. (2023) Computational prediction of disordered binding regions. *Comput. Struct. Biotechnol. J.*, **21**, 1487–1497.
13. Zhang,Y., Bao,W., Cao,Y., Cong,H., Chen,B. and Chen,Y. (2022) A survey on protein-DNA-binding sites in computational biology. *Brief. Funct. Genomics*, **21**, 357–375.
14. Jiang,Q., Jin,X., Lee,S.J. and Yao,S.W. (2017) Protein secondary structure prediction: a survey of the state of the art. *J. Mol. Graphics Model.*, **76**, 379–402.
15. Yan,J., Friedrich,S. and Kurgan,L. (2016) A comprehensive comparative review of sequence-based predictors of DNA- and RNA-binding residues. *Brief Bioinform.*, **17**, 88–105.
16. Miao,Z. and Westhof,E. (2015) A large-scale assessment of nucleic acids binding site prediction programs. *PLoS Comput. Biol.*, **11**, e1004639.
17. Si,J., Cui,J., Cheng,J. and Wu,R. (2015) Computational prediction of RNA-binding proteins and binding sites. *Int. J. Mol. Sci.*, **16**, 26303–26317.
18. Oldfield,C.J., Chen,K. and Kurgan,L. (2019) Computational prediction of secondary and supersecondary structures from protein sequences. *Methods Mol. Biol.*, **1958**, 73–100.
19. Zhang,J., Ma,Z. and Kurgan,L. (2019) Comprehensive review and empirical analysis of hallmarks of DNA-, RNA- and protein-binding residues in protein chains. *Brief Bioinform*, **20**, 1250–1268.
20. Wang,K., Hu,G., Wu,Z., Su,H., Yang,J. and Kurgan,L. (2020) Comprehensive survey and comparative assessment of RNA-binding residue predictions with analysis by RNA type. *Int. J. Mol. Sci.*, **21**, 6879.
21. Buchan,D.W., Minneci,F., Nugent,T.C., Bryson,K. and Jones,D.T. (2013) Scalable web services for the PSIPRED Protein Analysis Workbench. *Nucleic Acids Res.*, **41**, W349–W357.
22. Buchan,D.W.A. and Jones,D.T. (2019) The PSIPRED Protein Analysis Workbench: 20 years on. *Nucleic Acids Res.*, **47**, W402–W407.
23. Hou,J., Wu,T., Guo,Z., Quadir,F. and Cheng,J. (2020) The MULTICOM protein structure prediction server empowered by deep learning and contact distance prediction. *Methods Mol. Biol.*, **2165**, 13–26.
24. Cheng,J., Li,J., Wang,Z., Eickholt,J. and Deng,X. (2012) The MULTICOM toolbox for protein structure prediction. *BMC Bioinf.*, **13**, 65.
25. Barik,A., Katuwawala,A., Hanson,J., Paliwal,K., Zhou,Y. and Kurgan,L. (2020) DEPICTER: intrinsic disorder and disorder function prediction server. *J. Mol. Biol.*, **432**, 3379–3387.
26. Basu,S., Gsponer,J. and Kurgan,L. (2023) DEPICTER2: a comprehensive webserver for intrinsic disorder and disorder function prediction. *Nucleic Acids Res.*, **51**, W141–W147.
27. Olenyi,T., Marquet,C., Heinzinger,M., Kroger,B., Nikolova,T., Bernhofer,M., Sandig,P., Schutze,K., Littmann,M., Mirdita,M., *et al.* (2023) LambdaPP: fast and accessible protein-specific phenotype predictions. *Protein Sci.*, **32**, e4524.
28. Bernhofer,M., Dallago,C., Karl,T., Satagopam,V., Heinzinger,M., Littmann,M., Olenyi,T., Qiu,J., Schutze,K., Yachdav,G., *et al.* (2021) PredictProtein - predicting protein structure and function for 29 years. *Nucleic Acids Res.*, **49**, W535–W540.
29. Oates,M.E., Romero,P., Ishida,T., Ghalwash,M., Mizianty,M.J., Xue,B., Dosztanyi,Z., Uversky,V.N., Obradovic,Z., Kurgan,L., *et al.* (2013) D(2)P(2): database of disordered protein predictions. *Nucleic Acids Res.*, **41**, D508–D516.
30. Piovesan,D., Tabaro,F., Paladin,L., Necci,M., Micetic,I., Camilloni,C., Davey,N., Dosztanyi,Z., Meszaros,B., Monzon,A.M., *et al.* (2018) MobiDB 3.0: more annotations for intrinsic disorder, conformational diversity and interactions in proteins. *Nucleic Acids Res.*, **46**, D471–D476.
31. Piovesan,D., Necci,M., Escobedo,N., Monzon,A.M., Hatos,A., Micetic,I., Quaglia,F., Paladin,L., Ramasamy,P., Dosztanyi,Z., *et al.* (2021) MobiDB: intrinsically disordered proteins in 2021. *Nucleic Acids Res.*, **49**, D361–D367.
32. Piovesan,D., Del Conte,A., Clementel,D., Monzon,A.M., Bevilacqua,M., Aspromonte,M.C., Iserte,J.A., Orti,F.E., Marino-Buslje,C. and Tosatto,S.C.E. (2023) MobiDB: 10 years of intrinsically disordered proteins. *Nucleic Acids Res.*, **51**, D438–D444.
33. Zhao,B., Katuwawala,A., Oldfield,C.J., Dunker,A.K., Faraggi,E., Gsponer,J., Kloczkowski,A., Malhis,N., Mirdita,M., Obradovic,Z., *et al.* (2021) DescribePROT: database of amino acid-level protein structure and function predictions. *Nucleic Acids Res.*, **49**, D298–D308.
34. Zhou,T., Rong,J., Liu,Y., Gong,W. and Li,C. (2022) An ensemble approach to predict binding hotspots in protein-RNA interactions based on SMOTE data balancing and Random grouping feature selection strategies. *Bioinformatics*, **38**, 2452–2458.
35. Hou,C., Li,Y., Wang,M., Wu,H. and Li,T. (2022) Systematic prediction of degrons and E3 ubiquitin ligase binding via deep learning. *BMC Biol.*, **20**, 162.
36. Cermakova,K. and Hodges,H.C. (2023) Interaction modules that impart specificity to disordered protein. *Trends Biochem. Sci.*, **48**, 477–490.
37. Zhao,B., Katuwawala,A., Oldfield,C.J., Hu,G., Wu,Z., Uversky,V.N. and Kurgan,L. (2021) Intrinsic disorder in Human RNA-binding proteins. *J. Mol. Biol.*, **433**, 167229.
38. Tamburrini,K.C., Pesce,G., Nilsson,J., Gondelaud,F., Kajava,A.V., Berrin,J.G. and Longhi,S. (2022) Predicting protein conformational disorder and disordered binding sites. *Methods Mol. Biol.*, **2449**, 95–147.
39. Emonts,J. and Buyel,J.F. (2023) An overview of descriptors to capture protein properties-tools and perspectives in the context of QSAR modeling. *Comput. Struct. Biotechnol. J.*, **21**, 3234–3247.
40. Waury,K., Willemse,E.A.J., Vanmechelen,E., Zetterberg,H., Teunissen,C.E. and Abeln,S. (2022) Bioinformatics tools and data resources for assay development of fluid protein biomarkers. *Biomark. Res.*, **10**, 83.
41. Mackmull,M.T., Nagel,L., Sesterhenn,F., Muntel,J., Grossbach,J., Stalder,P., Bruderer,R., Reiter,L., van de Berg,W.D.J., de Souza,N., *et al.* (2022) Global, in situ analysis of the structural proteome in individuals with Parkinson's disease to identify a new class of biomarker. *Nat. Struct. Mol. Biol.*, **29**, 978–989.
42. Faraggi,E., Zhou,Y. and Kloczkowski,A. (2014) Accurate single-sequence prediction of solvent accessible surface area using local and global features. *Proteins*, **82**, 3170–3176.
43. Meng,F. and Kurgan,L. (2016) DFLpred: high-throughput prediction of disordered flexible linker regions in protein sequences. *Bioinformatics*, **32**, i341–i350.
44. Peng,Z. and Kurgan,L. (2015) High-throughput prediction of RNA, DNA and protein binding regions mediated by intrinsic disorder. *Nucleic Acids Res.*, **43**, e121.
45. Yan,J. and Kurgan,L. (2017) DRNApred, fast sequence-based method that accurately predicts and discriminates DNA- and RNA-binding residues. *Nucleic Acids Res.*, **45**, e84.
46. Mirdita,M., Steinegger,M. and Soding,J. (2019) MMseqs2 desktop and local web server app for fast, interactive sequence searches. *Bioinformatics*, **35**, 2856–2858.
47. Steinegger,M. and Soding,J. (2017) MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.*, **35**, 1026–1028.
48. Malhis,N., Jacobson,M. and Gsponer,J. (2016) MoRFchibi SYSTEM: software tools for the identification of MoRFs in protein sequences. *Nucleic Acids Res.*, **44**, W488–W493.

49. Zhang,J. and Kurgan,L. (2019) SCRIBER: accurate and partner type-specific prediction of protein-binding residues from proteins sequences. *Bioinformatics*, **35**, i343–i353.
50. Teufel,F., Almagro Armenteros,J.J., Johansen,A.R., Gislason,M.H., Pihl,S.I., Tsirigos,K.D., Winther,O., Brunak,S., von Heijne,G. and Nielsen,H. (2022) SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nat. Biotechnol.*, **40**, 1023–1025.
51. Peng,K., Radivojac,P., Vucetic,S., Dunker,A.K. and Obradovic,Z. (2006) Length-dependent prediction of protein intrinsic disorder. *BMC Bioinf.*, **7**, 208.
52. Hu,G., Katuwawala,A., Wang,K., Wu,Z., Ghadermarzi,S., Gao,J. and Kurgan,L. (2021) flDPnn: accurate intrinsic disorder prediction with putative propensities of disorder functions. *Nat. Commun.*, **12**, 4438.
53. Wang,D., Liu,D., Yuchi,J., He,F., Jiang,Y., Cai,S., Li,J. and Xu,D. (2020) MusiteDeep: a deep-learning based webserver for protein post-translational modification site prediction and visualization. *Nucleic Acids Res.*, **48**, W140–W146.
54. Wang,D., Zeng,S., Xu,C., Qiu,W., Liang,Y., Joshi,T. and Xu,D. (2017) MusiteDeep: a deep-learning framework for general and kinase-specific phosphorylation site prediction. *Bioinformatics*, **33**, 3909–3916.
55. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinf.*, **10**, 421.
56. Hu,G. and Kurgan,L. (2019) Sequence similarity searching. *Curr. Protoc. Protein. Sci.*, **95**, e71.
57. Pejaver,V., Hsu,W.L., Xin,F.X., Dunker,A.K., Uversky,V.N. and Radivojac,P. (2014) The structural and functional signatures of proteins that undergo multiple events of post-translational modification. *Protein Sci.*, **23**, 1077–1093.
58. Necci,M., Piovesan,D., Predictors,C., DisProt,C. and Tosatto,S.C.E. (2021) Critical assessment of protein intrinsic disorder prediction. *Nat. Methods*, **18**, 472–481.
59. Lang,B. and Babu,M.M. (2021) A community effort to bring structure to disorder. *Nat. Methods*, **18**, 454–455.
60. Zhao,B. and Kurgan,L. (2022) Deep learning in prediction of intrinsic disorder in proteins. *Comput. Struct. Biotechnol. J.*, **20**, 1286–1294.
61. Wang,K. and Samudrala,R. (2006) Incorporating background frequency improves entropy-based residue conservation measures. *BMC Bioinf.*, **7**, 385.
62. Styczynski,M.P., Jensen,K.L., Rigoutsos,I. and Stephanopoulos,G. (2008) BLOSUM62 miscalculations improve search performance. *Nat. Biotechnol.*, **26**, 274–275.
63. Suzek,B.E., Wang,Y., Huang,H., McGarvey,P.B., Wu,C.H. and UniProt,C. (2015) UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, **31**, 926–932.
64. Quaglia,F., Meszaros,B., Salladini,E., Hatos,A., Pancsa,R., Chemes,L.B., Pajkos,M., Lazar,T., Pena-Diaz,S., Santos,J., *et al.* (2022) DisProt in 2022: improved quality and accessibility of protein intrinsic disorder annotation. *Nucleic Acids Res.*, **50**, D480–D487.
65. Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
66. Tien,M.Z., Meyer,A.G., Sydykova,D.K., Spielman,S.J. and Wilke,C.O. (2013) Maximum allowed solvent accessibilites of residues in proteins. *PLoS One*, **8**, e80635.
67. Biro,B., Zhao,B. and Kurgan,L. (2022) Complementarity of the residue-level protein function and structure predictions in human proteins. *Comput. Struct. Biotechnol. J.*, **20**, 2223–2234.
68. Toufektchan,E. and Toledo,F. (2018) The Guardian of the Genome revisited: p53 downregulates genes required for telomere maintenance, DNA repair, and centromere structure. *Cancers (Basel)*, **10**, 135.
69. Ferreon,J.C., Lee,C.W., Arai,M., Martinez-Yamout,M.A., Dyson,H.J. and Wright,P.E. (2009) Cooperative regulation of p53 by modulation of ternary complex formation with CBP/p300 and HDM2. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 6591–6596.
70. Wells,M., Tidow,H., Rutherford,T.J., Markwick,P., Jensen,M.R., Mylonas,E., Svergun,D.I., Blackledge,M. and Fersht,A.R. (2008) Structure of tumor suppressor p53 and its intrinsically disordered N-terminal transactivation domain. *Proc. Natl. Acad. Sci. U.S.A.*, **105**, 5762–5767.
71. Oldfield,C.J., Meng,J., Yang,J.Y., Yang,M.Q., Uversky,V.N. and Dunker,A.K. (2008) Flexible nets: disorder and induced fit in the associations of p53 and 14-3-3 with their partners. *BMC Genomics*, **9**(Suppl. 1), S1.
72. Feng,H., Jenkins,L.M., Durell,S.R., Hayashi,R., Mazur,S.J., Cherry,S., Tropea,J.E., Miller,M., Wlodawer,A., Appella,E., *et al.* (2009) Structural basis for p300 Taz2-p53 TAD1 binding and modulation by phosphorylation. *Structure*, **17**, 202–210.
73. Mujtaba,S., He,Y., Zeng,L., Yan,S., Plotnikova,O.,,Sachchidanand, Sanchez,R., Zeleznik-Le,N.J., Ronai,Z. and Zhou,M.M. (2004) Structural mechanism of the bromodomain of the coactivator CBP in p53 transcriptional activation. *Mol. Cell*, **13**, 251–263.
74. Lidor Nili,E., Field,Y., Lubling,Y., Widom,J., Oren,M. and Segal,E. (2010) p53 binds preferentially to genomic regions with high DNA-encoded nucleosome occupancy. *Genome Res.*, **20**, 1361–1368.
75. McLure,K.G. and Lee,P.W. (1998) How p53 binds DNA as a tetramer. *EMBO J.*, **17**, 3342–3350.
76. Hornbeck,P.V., Kornhauser,J.M., Latham,V., Murray,B., Nandhikonda,V., Nord,A., Skrzypek,E., Wheeler,T., Zhang,B. and Gnad,F. (2019) 15 years of PhosphoSitePlus(R): integrating post-translationally modified sites, disease variants and isoforms. *Nucleic Acids Res.*, **47**, D433–D441.
77. Paysan-Lafosse,T., Blum,M., Chuguransky,S., Grego,T., Pinto,B.L., Salazar,G.A., Bileschi,M.L., Bork,P., Bridge,A., Colwell,L., *et al.* (2023) InterPro in 2022. *Nucleic Acids Res.*, **51**, D418–D427.
78. Zhang,C., Zhang,X., Freddolino,P.L. and Zhang,Y. (2023) BioLiP2: an updated structure database for biologically relevant ligand-protein interactions. *Nucleic Acids Res.* https://doi.org/10.1093/nar/gkad630.
79. Zhao,B., Ghadermarzi,S. and Kurgan,L. (2023) Comparative evaluation of AlphaFold2 and disorder predictors for prediction of intrinsic disorder, disorder content and fully disordered proteins. *Comput. Struct. Biotechnol. J.*, **21**, 3248–3258.
80. Wilson,C.J., Choy,W.Y. and Karttunen,M. (2022) AlphaFold2: a role for disordered protein/region prediction?*Int. J. Mol. Sci.*, **23**, 4591.

Received: September 14, 2023. **Revised:** October 12, 2023. **Editorial Decision:** October 13, 2023. **Accepted:** October 16, 2023