DisProt in 2022: improved quality and accessibility of protein intrinsic disorder annotation

Federica Quaglia [1,2], Bálint Mészáros [3], Edoardo Salladini [2], András Hatos [2], Rita Pancsa [4], Lucía B. Chemes [5], Mátyás Pajkos [6], Tamas Lazar [7,8], Samuel Peña-Díaz [9,10], Jaime Santos [9,10], Veronika Ács [4], Nazanin Farahi [7,8], Erzsébet Fichó [4,11], Maria Cristina Aspromonte [12,13], Claudio Bassot [14], Anastasia Chasapi [15], Norman E. Davey [16], Radoslav Davidović [17], Laszlo Dobson [3,4], Arne Elofsson [14], Gábor Erdős [6], Pascale Gaudet [18], Michelle Giglio [19], Juliana Glavina [5], Javier Iserte [20], Valentín Iglesias [9,10], Zsófia Kálmán [21], Matteo Lambrughi [22], Emanuela Leonardi [12,13], Sonia Longhi [23], Sandra Macedo-Ribeiro [24,25], Emiliano Maiani [22], Julia Marchetti [26], Cristina Marino-Buslje [20], Attila Mészáros [7,8], Alexander Miguel Monzon [2], Giovanni Minervini [2], Suvarna Nadendla [19], Juliet F. Nilsson [23], Marian Novotný [27], Christos A. Ouzounis [15,28], Nicolás Palopoli [26], Elena Papaleo [22,29], Pedro José Barbosa Pereira [24,25], Gabriele Pozzati [14], Vasilis J Promponas [30], Jordi Pujols [9,10], Alma Carolina Sanchez Rocha [31], Martin Salas [26], Luciana Rodriguez Sawicki [26], Eva Schad [4], Aditi Shenoy [14], Tamás Szaniszló [6], Konstantinos D. Tsirigos [32], Nevena Veljkovic [17], Gustavo Parisi [26], Salvador Ventura [9,10,33], Zsuzsanna Dosztányi [6], Peter Tompa [4,7,8], Silvio C.E. Tosatto [2,*] and Damiano Piovesan [2,*]

[1]Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies, National Research Council (CNR-IBIOM), Bari, Italy, [2]Department of Biomedical Sciences, University of Padova, Padova, Italy, [3]Structural and Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg 69117, Germany, [4]Institute of Enzymology, Research Centre for Natural Sciences, Budapest 1117, Hungary, [5]Instituto de Investigaciones Biotecnológicas (IIBiO-CONICET), Universidad Nacional de San Martín, Av. 25 de Mayo y Francia, CP1650 Buenos Aires, Argentina, [6]Department of Biochemistry, Eötvös Loránd University, Pázmány Péter stny 1/c, Budapest H-1117, Hungary, [7]VIB-VUB Center for Structural Biology, Vlaams Instituut voor Biotechnologie, Brussels, Belgium, [8]Structural Biology Brussels (SBB), Bioengineering Sciences Department, Vrije Universiteit Brussel (VUB), Brussels, Belgium, [9]Institut de Biotecnologia i Biomedicina, Universitat Autònoma de Barcelona, Barcelona, Spain, [10]Departament de Bioquímica i Biologia Molecular, Universitat Autònoma de Barcelona, Barcelona, Spain, [11]Cytocast Kft., Vecsés, Hungary, [12]Department of Woman and Child Health, University of Padova, Padova, Italy, [13]Pediatric Research Institute, Città della Speranza, Padova, Italy, [14]Science for Life Laboratory, Department of Biochemistry and Biophysics, Stockholm University, 171 21 Solna, Sweden, [15]Biological Computation & Process Laboratory, Chemical Process & Energy Resources Institute, Centre for Research & Technology Hellas, Thermi, Thessalonica 57001, Greece, [16]Institute of Cancer Research, Chester Beatty Laboratories, 237 Fulham Rd, Chelsea, London, UK, [17]Laboratory for Bioinformatics and Computational Chemistry, Vinča Institute of Nuclear Sciences, National Institute of the Republic of Serbia, University of Belgrade, 11000 Belgrade, Serbia, [18]Swiss-Prot group, SIB Swiss Institute of Bioinformatics, Geneva, Switzerland, [19]Institute for Genome Sciences, University of Maryland School of Medicine 670 W. Baltimore St., Baltimore, MD 21201, USA, [20]Bioinformatics Unit, Fundación Instituto Leloir, Buenos Aires,

*To whom correspondence should be addressed. Tel: +39 049 827 6269; Email: damiano.piovesan@unipd.it
Correspondence may also be addressed to Silvio C.E. Tosatto. Email: silvio.tosatto@unipd.it

C1405BWE, Argentina, [21]Faculty of Information Technology and Bionics, Pázmány Péter Catholic University, Práter u. 50/A, 1083 Budapest, Hungary, [22]Cancer Structural Biology, Danish Cancer Society Research Center, Strandboulevarden 49, 2100 Copenhagen, Denmark, [23]Lab. Architecture et Fonction des Macromolécules Biologiques (AFMB), UMR 7257, Aix Marseille University and Centre National de la Recherche Scientifique (CNRS), 163 Avenue de Luminy, Case 932, 13288, Marseille, France, [24]Instituto de Biologia Molecular e Celular (IBMC), Universidade do Porto, 4200-135 Porto, Portugal, [25]Instituto de Investigação e Inovação em Saúde (i3S), Universidade do Porto, 4200-135 Porto, Portugal, [26]Departamento de Ciencia y Tecnología, Universidad Nacional de Quilmes - CONICET, Bernal, Buenos Aires B1876BXD, Argentina, [27]Dep. of Cell Biology, Faculty of Science, Vinicna 7, 128 43, Prague, Czech Republic, [28]Biological Computation & Computational Biology Group, Artificial Intelligence & Information Analysis Lab, Department of Computer Science, Aristotle University of Thessalonica, Thessalonica 54124, Greece, [29]Cancer Systems Biology, Section for Bioinformatics, Department of Health and Technology, Technical University of Denmark, Lyngby, Denmark, [30]Bioinformatics Research Laboratory, Department of Biological Sciences, University of Cyprus, Nicosia, Cyprus, [31]Department of Cell Biology, Faculty of Science, Charles University, BIOCEV, Prague , Czech Republic, [32]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, UK and [33]ICREA, Barcelona, Spain

ABSTRACT

The Database of Intrinsically Disordered Proteins (DisProt, URL: https://disprot.org) is the major repository of manually curated annotations of intrinsically disordered proteins and regions from the literature. We report here recent updates of DisProt version 9, including a restyled web interface, refactored Intrinsically Disordered Proteins Ontology (IDPO), improvements in the curation process and significant content growth of around 30%. Higher quality and consistency of annotations is provided by a newly implemented reviewing process and training of curators. The increased curation capacity is fostered by the integration of DisProt with APICURON, a dedicated resource for the proper attribution and recognition of biocuration efforts. Better interoperability is provided through the adoption of the Minimum Information About Disorder (MIADE) standard, an active collaboration with the Gene Ontology (GO) and Evidence and Conclusion Ontology (ECO) consortia and the support of the ELIXIR infrastructure.

INTRODUCTION

Whereas our traditional view of protein function is rooted in the model of proteins assuming a stable structure, a well-defined 3D fold, it is now >20 years since the concept of structural disorder of proteins has been proposed (1,2). The existence and functional importance of intrinsically disordered proteins/regions (IDPs/IDRs) is now generally accepted (3), with >1500 PubMed publications mentioning disordered proteins every year.

The prediction of protein disorder from sequence, for example, has always been an area of continuous activity. Recently, it has received a boost with the establishment of the Critical Assessment of Protein Intrinsic Disorder prediction (CAID) experiment as a community-wide blind test to compare state-of-the-art approaches to predict disorder (4).

As new disorder prediction methods keep emerging (5–7), CAID takes on monitoring the field in real time, aiming to establish dependable standards. This ambition has a special caveat, as predicting and identifying regions in IDPs/IDRs that engage in functional interactions remains a significant challenge (6,8).

The prediction of such functional regions may support and inspire dedicated experimental approaches. The Eukaryotic Linear Motif (ELM) database is a primary repository of such data (9). Of similar ambition pursued along different lines, the database of fuzzy complexes, FuzDB, compiles experimentally observed fuzzy protein complexes, in which intrinsic disorder is maintained upon partner interaction, directly impacting biological function (10). An instructive example of this behavior is the extremely tight and functional interaction between the disordered histone H1 and its chaperone prothymosin-α, which retain their highly dynamic, fully disordered state in that complex (11). Structural interpretation of such a behavior may be assisted by data in the Protein Ensemble Database (PED) that encompasses experimentally determined structural ensembles of IDPs/IDRs (12).

There appears to be a consensus that these and other types of data and approaches will act in synergy to drive the field forward (13), toward reaching a better structural-functional understanding of the 'disorderome'. Combined with data in the MobiDB database, which provides predictions and annotations for all IDPs/IDRs (14), this effort is also critical for integrating disorder-related information into other data resources, such as UniProtKB (15) or PDBe (16).

Not surprisingly, the field is also getting strong impetus from traditional structural biology: as missing regions from solved structures constitute a good proxy for structural disorder, improvements in structure determination techniques boost the identification of structural disorder. Cryo-electron microscopy has now advanced to a state that yields structures at atomic resolution (17). Similarly, the recent success of computational structure prediction by AlphaFold 2 (18) in overtaking the Critical Assessment

of Structure Prediction competition, CASP14 (19), cannot be dismissed. AlphaFold's 2-based structure predictions of the entire human proteome (20) resulted in 'only' 58% of residues confidently covered with predicted structures, suggesting that structural disorder may be even more prevalent in proteomes than previously thought (21). Whereas it remains to be confirmed if certain AlphaFold 2 metrics can be harnessed for disorder prediction (20), it is without doubt that these achievements will help us navigate further through the 'dark proteome' (22).

Disordered proteins are often involved in disease, yet they represent a largely unexplored target for drug development (23,24). Recent successes in the field, i.e. targeting c-Myc (25), androgen receptor (26) or alpha-synuclein (27), encourage us to adhere to this ambition. The concept of liquid-liquid phase separation (LLPS) leading to the formation of membraneless organelles (MLOs) raises hope that IDP/IDR function and therapeutic targeting can now be approached from a novel angle (28,29).

To reflect the steady progress of the protein disorder field, it is important to update and upgrade DisProt as one of the primary resources of manually curated, experimentally confirmed protein disorder. The previous release of the database, DisProt 8 (30) contained about 1500 entries and 3500 disordered protein regions. In the current release, DisProt 9, not only did we increase these numbers, but also improved the reliability of entries by introducing a reviewing process. In addition, a great effort was allocated to training activities by providing DisProt biocurators with detailed curation guidelines and virtual training sessions and a published protocol describing how to explore manually curated annotations in DisProt (13), alongside the dissemination of new database content in a dedicated DisProt blog (https://disprot.org/blog/) and Twitter account (https://twitter.com/disprot_db). DisProt 9 presents a new graphical interface and updated features, such as the integration of two ontologies and the connection with API-CURON, a database to credit and acknowledge the work of biocurators (31). With these improvements, DisProt continues to be a primary resource of protein disorder for the structural-molecular biology community.

PROGRESS AND NEW FEATURES

Database content

DisProt 9 includes 2038 protein entries and 4,477 pieces of evidence of state transitions, interactions and functions, featuring a 30% increase over the last release (DisProt 8), along with 2578 publications annotated, accounting for a 28% increase compared with DisProt 8. In addition, about 14% of the annotations, corresponding to 237 entries, have been reviewed and validated by an expert biocurator, e.g. the High mobility group protein HMG-I/HMG-Y (DP00040) has been thoroughly reannotated by revising spurious annotations and integrating new functions, such as *DNA bending* (GO:0008301, IDPO:00514), *DNA binding* (GO:0003677, IDPO:00065), *RNA binding (*GO:0003723, IDPO:00066) and *protein binding* (GO:0005515, IDPO:00063).

In DisProt 9, annotations of amino acid repeats displaying the typical properties of IDPs, specifically poly-glutamate (polyE), poly-lysine (polyK) and poly-arginine (polyR) regions, have been added based on indirect evidence. These regions are always disordered but experimental evidence is provided for only a handful of cases or for engineered fragments. For example the polyE repeat in the PEVK region of human titin has a Stokes radius 2–4 times larger than expected based on its molecular mass, and it shows a minimum at 200 nm in its circular dichroism (CD) spectrum (32). In order to be included in DisProt a repeated region must be predicted by MobiDB-lite (33) (and be available in MobiDB) as a negative or positive polyelectrolyte subregion (14,32) and be at least 10 residues long. In DisProt 9, such cases are highlighted by a specific evidence code, *curator inference from database* (ECO:0007636).

The distribution of regions based on the experimental detection method is shown in Figure 1. In accordance with the recent improvements that recognize cryo-EM as a well-established technique in structural biology (17), DisProt includes a total of 149 IDRs annotated by this method.

Figure 2 shows the length distribution of regions annotated using a specific experimental technique. Atomic resolution techniques such as X-ray (ECO:0006187) and NMR (ECO:0006252) dominate the experiments used to characterize short (<100 residues) disordered regions, while other complementary methods and far-UV CD (ECO:0006179) are the techniques mostly used in detecting and characterizing longer IDRs.

Figure 3 illustrates the amino acid composition and fold increase (enrichment) of DisProt proteins, as compared to TrEMBL distribution (release 2021_03). The fold increase, which is the difference between DisProt and TrEMBL frequencies normalized by TrEMBL, highlights amino acids which are over- and under-represented in DisProt IDRs. DisProt regions are enriched in disorder-promoting residues (Q, K, P, E, S), mostly charged and hydrophilic, while the hydrophobic, order-promoting residues (F, A, V, I, L), as well as arginine (R) are depleted and aromatic residues (W, Y, F) are strongly depleted. This amino acid distribution is in line with previously published results (34,35).

Thematic datasets

Starting 2020, DisProt periodically releases '*thematic datasets*' describing biological areas where IDPs play a crucial role. The first two thematic DisProt datasets illustrate the role of intrinsic disorder in *unicellular toxins and antitoxins* (December 2020) and *extracellular matrix proteins* (June 2021), by collecting carefully curated examples of IDPs and IDRs involved in the toxin-antitoxin system and in the extracellular matrix composition and function, respectively. DisProt thematic datasets are easily accessible from the DisProt home page under the '*Datasets*' section. All the entries included in these datasets are tagged with the name of the *theme* and it is possible to download them as single files.

Ontology

In DisProt 9, the previous Disordered Ontology (DO) has been renamed Intrinsically Disordered Protein Ontology
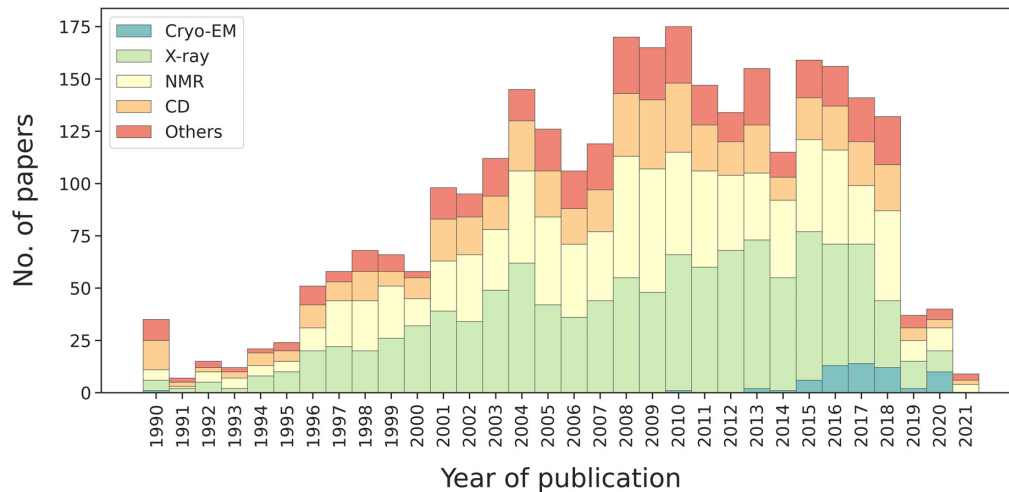
Figure 1. Number of experimental evidence of disorder used in DisProt by year of publication. Different colors correspond to different experimental techniques as reported in the corresponding publications. Publications older than 1990 are grouped in the first bar.
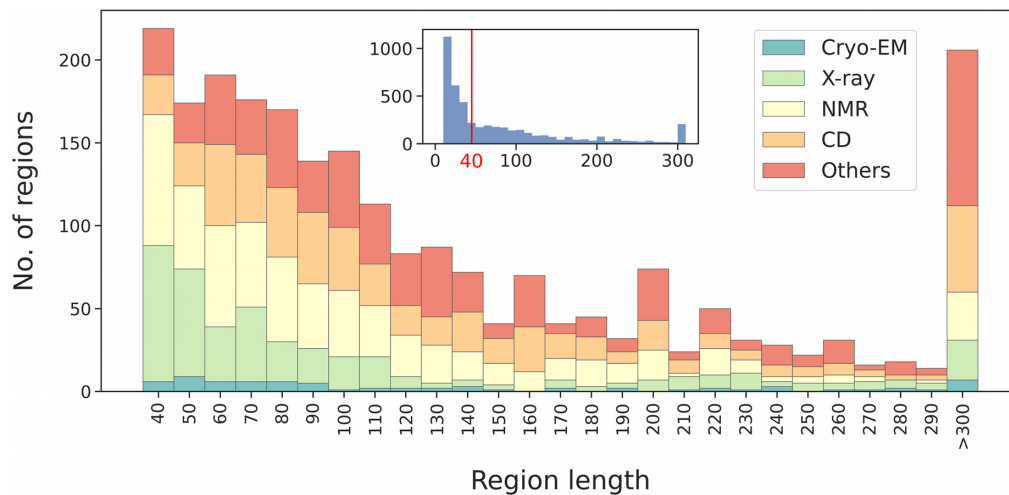
Figure 2. Region length distribution of pieces of evidence inferred using various experimental techniques in DisProt. Data in the main plot starts at length 40, the full distribution is shown in the inset plot.

(IDPO) and has been completely refactored. The 'experimental method' branch has been mapped one-to-one to a subset of the Evidence and Conclusion Ontology (ECO) (36), which was recently updated with new disorder-specific terms, e.g. IDPO:00125 maps to the corresponding *small-angle X-ray scattering* (ECO:0006182) ECO term. In addition, several new non-IDP-specific terms were introduced to cover a plethora of missing techniques, e.g. bait-prey protein pull-down evidence (ECO:0006249) is widely used to assess interactions and functions of IDPs and IDRs.

Similarly, all interaction terms and some function terms have been mapped to the Gene Ontology (GO) (37–39). Diverse modes of action of IDPs have been completely revised to provide the highest possible level of detail in the characterization of disorder functions. Despite GO being designed to annotate the whole protein, several of its terms are well suited to describe IDR behaviour, in particular those in the molecular function branch. IDPO terms were merged, mapped to existing GO terms or directly integrated into

GO, thanks to an active collaboration between the DisProt and GO consortia.

All active functions, which exert their effect via molecular recognition on another interactor, are now annotated using GO terms, while self-regulatory functions (self-regulation or self-assembly via protein interactions in *cis*) and functions directly arising from the disordered state (*entropic chains*) are annotated by specific IDPO terms since corresponding terms are not yet available in GO. *Entropic bristle*, *entropic spring* and *entropic clock* terms have now been merged under the *entropic chain* term. The *Flexible N-terminal tail* and *flexible C-terminal tail* terms were introduced to highlight the presence of disorder in protein terminal regions.

The DisProt 9 website incorporates a new dedicated page for each IDPO term, e.g. https://disprot.org/idpo/IDPO:00501 for the *entropic chain* term. IDPO term pages include the identifier, name and definition of the term, its relationships with other terms and, when available, cross-references
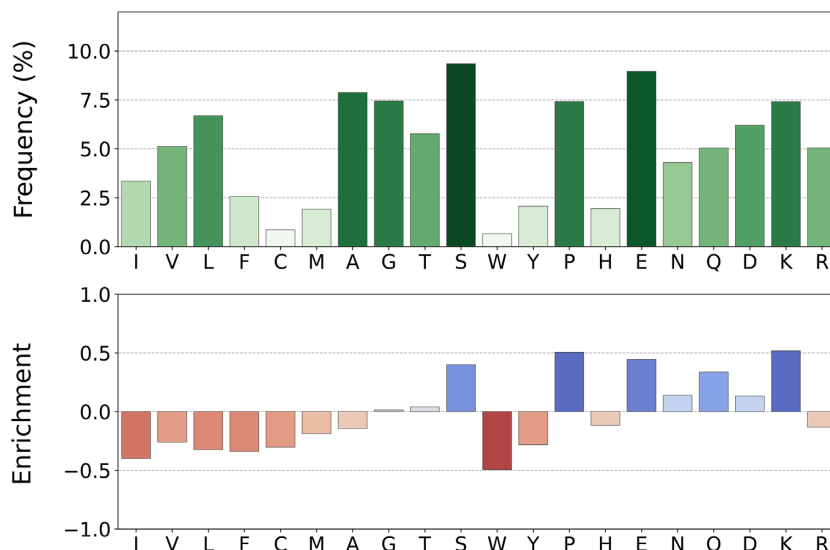
Figure 3. Amino acid composition of DisProt regions. Amino acids are sorted by the Kyte-Doolittle hydrophobicity scale. The amino acid frequency is calculated considering only disordered residues. The enrichment is calculated and normalized over the TrEMBL database frequencies (release 2021_03). Color intensity is proportional to bar height.

to external ontologies, e.g. Gene Ontology. Moreover, they list all entries annotated with that term.

The adoption of stable and well established ontologies, GO and ECO, plays a crucial role by allowing the curators to effortlessly expand the coverage of functions and experimental techniques available in DisProt. Since its partnership with the GO and ECO consortia, DisProt has become active in the definition of new terms that fulfill ontology-specific rules and constraints. Requests for new terms coming from the IDP community are welcome but must undergo GO and ECO approval. In the long run, all terms used by DisProt will be mapped to well established ontologies.

Experimental ambiguity and MIADE

A number of DisProt annotations come from experiments performed under extreme conditions. These annotations, which were tagged as ambiguous in the previous DisProt version, are now described using the Minimum Information About Disorder Experiments (MIADE) guidelines (https://www.psidev.info/groups/intrinsically-disordered-proteins) (40). MIADE defines the minimal fundamental parameters that unambiguously characterize a disorder-related experiment. With MIADE, it is possible to evaluate and compare experimental evidence coming from other resources adopting the same standard objectively.

In DisProt 9, MIADE is implemented by including the following annotation fields: (i) *sequence construct* features, e.g. the exact sequence from the experiment along with its modifications, such as PTMs and mutations, (ii) the *experimental conditions*, i.e. pH, temperature, pressure and redox potential and (iii) the *experimental components*, e.g. small-molecules and membranes, along with values, deviation and additional variables. Further specifications are provided in the help page on the DisProt website.

Biocuration and APICURON

DisProt annotations are provided by both professional and community biocurators. The whole community of biocurators is supported by a team of senior biocurators who check and validate annotations and deliver training material including face-to-face activities and a detailed curation manual. In order to properly attribute the curation effort, DisProt is now connected to APICURON (31). APICURON collects and tracks biocuration events from manually curated resources and implements gamification concepts, i.e. badges, medals and leaderboards, to promote biocurator engagement. The DisProt website integrates an APICURON widget (https://disprot.org/release-notes) which provides the ranking, number of activities and scores of the curators. In APICURON, all DisProt biocurators have a dedicated profile page that groups all the achieved badges and medals, along with their curation activity.

Additional annotations can be provided directly by the DisProt users through a 'Contact us' page. In DisProt 9, along with the '*Leave a comment*' tab for feedback on the site experience, a new '*Submit a new annotation*' section has been added.

Implementation

DisProt 9 adopts the Document Versioning Pattern to precisely track all changes between different versions of the same record. This improves reproducibility, simplifies tracking of curation activities and allows the user to compare different versions of an annotated entry directly on the website. Each protein has a *history* page and each IDR has an assigned version number. Whenever a curator modifies an existing region, the version counter increases by one.

DisProt exposes disordered region descriptions with Bioschemas metadata, including disordered region

boundaries and type of disorder. The information content from the entry pages is aggregated together with complementary resources (MobiDB and PED) to the IDPcentral registry and knowledgebase (41).

The DisProt web application source code underwent a major upgrade, updating Angular version 12 and transitioning to Bootstrap version 5. The DisProt interface is now more intuitive, thanks to a minimalistic, consistent design between different components. The upgraded curation interface allows for more efficient work from biocurators, while raising the standards of curated data quality thanks to real-time syntax checks and cross-validation with various third-party data resources.

Outreach activities

DisProt has an active social media presence on Twitter (https://twitter.com/disprot_db). This account promotes communication with users and experts in the field. It provides updated information about the latest release, webinars and training activities. Finally, a dedicated blog website (https://disprot.org/blog) contains descriptive extended posts about thematic datasets, general statistics and novel, interesting annotations. Training material for the users is provided via a published protocol describing how to explore manually curated annotations in DisProt (13). Training material for the curators is offered via recorded webinars and an updated user manual.

CONCLUSIONS AND FUTURE WORK

DisProt is the gold standard for IDP/IDR annotations, and serves the community as a fundamental resource that drives biological hypotheses, experimental design and the training and benchmarking of disorder and function predictors. Compared to the previous version, it has improved data accessibility and quality, and significantly increased annotation volume. The content is updated frequently and is now more focused on function. A team of expert reviewers validate annotations provided by the community and continuously check the literature for novel experimental evidence.

DisProt is well connected to other databases and consortia, and is active in the development of new standards and ontologies. The *Intrinsically Disordered Proteins Ontology* (IDPO) has been refactored and systematically cross-referenced with *Gene Ontology* (GO) and *Evidence and Conclusion Ontology* (ECO). Experimental setup is now captured in an unambiguous and structured way by implementing the *Minimum Information About Disorder* (MIADE) standard.

The DisProt technological infrastructure has been renewed to improve reproducibility by implementing exhaustive versioning of all entries. The community of curators is engaged, structured, well trained and continuously updated. Finally, DisProt is now connected with APICURON to provide live tracking and proper attribution of the curation effort.

DisProt is committed to the reduction of false negative annotations, resolution of inconsistent annotations (e.g. between close homologs) and growth of functional descriptions. At the consortium level, DisProt is active on the def-

inition of MIADE guidelines, creation of a standard controlled vocabulary to define conformational states (in collaboration with the structural biology community). From a technological point of view, DisProt is working to meet the format standards to export its annotations to core data resources such as GO, PDB, IntAct and UniProt. The long-term maintenance of DisProt is guaranteed by its central role within the European Union's Horizon 2020 IDPfun program and the ELIXIR IDP Community, the reference scientific communities involved in the study of intrinsically disordered proteins.

DATA AVAILABILITY

The data that support the findings of this study are openly available in DisProt at https://disprot.org/.

REFERENCES

1. Romero,P., Obradovic,Z., Kissinger,C.R., Villafranca,J.E., Garner,E., Guilliot,S. and Dunker,A.K. (1998) Thousands of proteins likely to have long disordered regions. *Pac. Symp. Biocomput. Pac. Symp. Biocomput.*, **1998**, 437–448.

2. Wright,P.E. and Dyson,H.J. (2015) Intrinsically disordered proteins in cellular signalling and regulation. *Nat. Rev. Mol. Cell Biol.*, **16**, 18–29.

3. van der Lee,R., Buljan,M., Lang,B., Weatheritt,R.J., Daughdrill,G.W., Dunker,A.K., Fuxreiter,M., Gough,J., Gsponer,J., Jones,D.T. *et al.* (2014) Classification of intrinsically disordered regions and proteins. *Chem. Rev.*, **114**, 6589–6631.

4. Necci,M., Piovesan,D., Predictors,C., Curators,D.P. and Tosatto,S.C.E. (2021) Critical assessment of protein intrinsic disorder prediction. *Nat. Methods*, **18**, 472–481.

5. 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.

6. Hanson,J., Paliwal,K.K., Litfin,T. and Zhou,Y. (2019) SPOT-Disorder2: improved protein intrinsic disorder prediction by ensembled deep learning. *Genomics Proteomics Bioinformatics*, **17**, 645–656.

7. Oberti,M. and Vaisman,I.I. (2020) cnnAlpha: protein disordered regions prediction by reduced amino acid alphabets and convolutional neural networks. *Proteins*, **88**, 1472–1481.

8. Mészáros,B., Erdős,G. and Dosztányi,Z. (2018) IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic. Acids. Res.*, **46**, W329–W337.

9. Kumar,M., Gouw,M., Michael,S., Sámano-Sánchez,H., Pancsa,R., Glavina,J., Diakogianni,A., Valverde,J.A., Bukirova,D., Čalyševa,J. *et al.* (2020) ELM—the eukaryotic linear motif resource in 2020. *Nucleic. Acids. Res.*, **48**, D296–D306.

10. Miskei,M., Antal,C. and Fuxreiter,M. (2017) FuzDB: database of fuzzy complexes, a tool to develop stochastic structure-function relationships for protein complexes and higher-order assemblies. *Nucleic Acids Res.*, **45**, D228–D235.

11. Borgia,A., Borgia,M.B., Bugge,K., Kissling,V.M., Heidarsson,P.O., Fernandes,C.B., Sottini,A., Soranno,A., Buholzer,K.J., Nettels,D. *et al.* (2018) Extreme disorder in an ultrahigh-affinity protein complex. *Nature*, **555**, 61–66.

12. Lazar,T., Martínez-Pérez,E., Quaglia,F., Hatos,A., Chemes,L.B., Iserte,J.A., Méndez,N.A., Garrone,N.A., Saldaño,T.E., Marchetti,J. *et al.* (2021) PED in 2021: a major update of the protein ensemble database for intrinsically disordered proteins. *Nucleic Acids Res.*, **49**, D404–D411.

13. Quaglia,F., Hatos,A., Piovesan,D. and Tosatto,S.C.E. (2020) Exploring manually curated annotations of intrinsically disordered proteins with disprot. *Curr. Protoc. Bioinforma.*, **72**, e107.

14. Piovesan,D., Necci,M., Escobedo,N., Monzon,A.M., Hatos,A., Mičetić,I., Quaglia,F., Paladin,L., Ramasamy,P., Dosztányi,Z. *et al.* (2021) MobiDB: intrinsically disordered proteins in 2021. *Nucleic Acids Res.*, **49**, D361–D367.

15. UniProt Consortium (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.*, **47**, D506–D515.

16. PDBe-KB consortium (2020) PDBe-KB: a community-driven resource for structural and functional annotations. *Nucleic Acids Res.*, **48**, D344–D353.

17. Nakane,T., Kotecha,A., Sente,A., McMullan,G., Masiulis,S., Brown,P.M.G.E., Grigoras,I.T., Malinauskaite,L., Malinauskas,T., Miehling,J. *et al.* (2020) Single-particle cryo-EM at atomic resolution. *Nature*, **587**, 152–156.

18. Jumper,J., Evans,R., Pritzel,A., Green,T., Figurnov,M., Ronneberger,O., Tunyasuvunakool,K., Bates,R., Žídek,A., Potapenko,A. *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**, 583–589.

19. Pereira,J., Simpkin,A.J., Hartmann,M.D., Rigden,D.J., Keegan,R.M. and Lupas,A.N. (2021) High-accuracy protein structure prediction in CASP14. *Proteins Struct. Funct. Bioinforma.*, https://doi.org/10.1002/prot.26171.

20. Tunyasuvunakool,K., Adler,J., Wu,Z., Green,T., Zielinski,M., Žídek,A., Bridgland,A., Cowie,A., Meyer,C., Laydon,A. *et al.* (2021) Highly accurate protein structure prediction for the human proteome. *Nature*, **596**, 590–596.

21. Pancsa,R. and Tompa,P. (2012) Structural disorder in eukaryotes. *PLoS One*, **7**, e34687.

22. Bhowmick,A., Brookes,D.H., Yost,S.R., Dyson,H.J., Forman-Kay,J.D., Gunter,D., Head-Gordon,M., Hura,G.L., Pande,V.S., Wemmer,D.E. *et al.* (2016) Finding our way in the dark proteome. *J. Am. Chem. Soc.*, **138**, 9730–9742.

23. Ruan,H., Sun,Q., Zhang,W., Liu,Y. and Lai,L. (2019) Targeting intrinsically disordered proteins at the edge of chaos. *Drug Discov. Today*, **24**, 217–227.

24. Santofimia-Castaño,P., Rizzuti,B., Xia,Y., Abian,O., Peng,L., Velázquez-Campoy,A., Neira,J.L. and Iovanna,J. (2020) Targeting intrinsically disordered proteins involved in cancer. *Cell. Mol. Life Sci. CMLS*, **77**, 1695–1707.

25. Boike,L., Cioffi,A.G., Majewski,F.C., Co,J., Henning,N.J., Jones,M.D., Liu,G., McKenna,J.M., Tallarico,J.A., Schirle,M. *et al.* (2021) Discovery of a functional covalent ligand targeting an intrinsically disordered cysteine within MYC. *Cell Chem. Biol.*, **28**, 4–13.

26. Ban,F., Leblanc,E., Cavga,A.D., Huang,C.-F., Flory,M.R., Zhang,F., Chang,M.E.K., Morin,H., Lallous,N., Singh,K. *et al.* (2021) Development of an androgen receptor inhibitor targeting the N-Terminal domain of androgen receptor for treatment of castration resistant prostate cancer. *Cancers*, **13**, 3488.

27. Tóth,G., Neumann,T., Berthet,A., Masliah,E., Spencer,B., Tao,J., Jobling,M.F., Gardai,S.J., Bertoncini,C.W., Cremades,N. *et al.* (2019) Novel small molecules targeting the intrinsically disordered structural ensemble of α-synuclein protect against diverse α-synuclein mediated dysfunctions. *Sci. Rep.*, **9**, 16947.

28. Mathieu,C., Pappu,R.V. and Taylor,J.P. (2020) Beyond aggregation: pathological phase transitions in neurodegenerative disease. *Science*, **370**, 56–60.

29. Cai,D., Liu,Z. and Lippincott-Schwartz,J. (2021) Biomolecular condensates and their links to cancer progression. *Trends Biochem. Sci.*, **46**, 535–549.

30. Hatos,A., Hajdu-Soltész,B., Monzon,A.M., Palopoli,N., Álvarez,L., Aykac-Fas,B., Bassot,C., Benítez,G.I., Bevilacqua,M., Chasapi,A. *et al.* (2020) DisProt: intrinsic protein disorder annotation in 2020. *Nucleic Acids Res.*, **48**, D269–D276.

31. Hatos,A., Quaglia,F., Piovesan,D. and Tosatto,S.C.E. (2021) APICURON: a database to credit and acknowledge the work of biocurators. *Database J. Biol. Databases Curation*, **2021**, baab019.

32. Duan,Y., DeKeyser,J.G., Damodaran,S. and Greaser,M.L. (2006) Studies on titin PEVK peptides and their interaction. *Arch. Biochem. Biophys.*, **454**, 16–25.

33. Necci,M., Piovesan,D., Clementel,D., Dosztányi,Z. and Tosatto,S.C.E. (2020) MobiDB-lite 3.0: fast consensus annotation of intrinsic disorder flavors in proteins. *Bioinformatics*, **36**, 5533–5534.

34. Dunker,A.K., Lawson,J.D., Brown,C.J., Williams,R.M., Romero,P., Oh,J.S., Oldfield,C.J., Campen,A.M., Ratliff,C.M., Hipps,K.W. *et al.* (2001) Intrinsically disordered protein. *J. Mol. Graph. Model.*, **19**, 26–59.

35. Necci,M., Piovesan,D. and Tosatto,S.C.E. (2018) Where differences resemble: sequence-feature analysis in curated databases of intrinsically disordered proteins. *Database J. Biol. Databases Curation*, **2018**, bay127.

36. Giglio,M., Tauber,R., Nadendla,S., Munro,J., Olley,D., Ball,S., Mitraka,E., Schriml,L.M., Gaudet,P., Hobbs,E.T. *et al.* (2019) ECO, the Evidence & Conclusion Ontology: community standard for evidence information. *Nucleic Acids Res.*, **47**, D1186–D1194.

37. Gene Ontology Consortium (2021) The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res.*, **49**, D325–D334.

38. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

39. Binns,D., Dimmer,E., Huntley,R., Barrell,D., O'Donovan,C. and Apweiler,R. (2009) QuickGO: a web-based tool for Gene Ontology searching. *Bioinforma. Oxf. Engl.*, **25**, 3045–3046.

40. Davey,N.E., Babu,M.M., Blackledge,M., Bridge,A., Capella-Gutierrez,S., Dosztanyi,Z., Drysdale,R., Edwards,R.J., Elofsson,A., Felli,I.C. *et al.* (2019) An intrinsically disordered proteins community for ELIXIR. *F1000Research*, **8**, 1753.

41. Gray,A.J.G., Papadopoulos,P., Mičetić,I. and Hatos,A. (2021) Exploiting bioschemas markup to populate IDPcentral. *BioHackrXiv*, https://doi.org/10.37044/osf.io/v3jct.