DisProt 7.0: a major update of the database of disordered proteins

Damiano Piovesan, Francesco Tabaro, Ivan Mičetić, Marco Necci, Federica Quaglia, Christopher J. Oldfield, Maria Cristina Aspromonte, Norman E. Davey, Radoslav Davidović, Zsuzsanna Dosztányi, et al.

Published online 24 November 2016 *Nucleic Acids Research, 2017, Vol. 45, Database issue* ***D219–D227***
doi: 10.1093/nar/gkw1056

DisProt 7.0: a major update of the database of disordered proteins

Damiano Piovesan[1,†], Francesco Tabaro[1,2,†], Ivan Mičetić[1], Marco Necci[1], Federica Quaglia[1], Christopher J. Oldfield[3], Maria Cristina Aspromonte[4], Norman E. Davey[5,6], Radoslav Davidović[7], Zsuzsanna Dosztányi[8,9], Arne Elofsson[10], Alessandra Gasparini[4], András Hatos[1,9], Andrey V. Kajava[11,12,13], Lajos Kalmar[9,14], Emanuela Leonardi[4], Tamas Lazar[15,16], Sandra Macedo-Ribeiro[17], Mauricio Macossay-Castillo[15,16], Attila Meszaros[9], Giovanni Minervini[1], Nikoletta Murvai[9], Jordi Pujols[18], Daniel B. Roche[11,12], Edoardo Salladini[19], Eva Schad[9], Antoine Schramm[19], Beata Szabo[9], Agnes Tantos[9], Fiorella Tonello[1,20], Konstantinos D. Tsirigos[10], Nevena Veljković[7], Salvador Ventura[18], Wim Vranken[15,16,21], Per Warholm[10], Vladimir N. Uversky[22,23], A. Keith Dunker[3], Sonia Longhi[19,*], Peter Tompa[9,15,16,*] and Silvio C.E. Tosatto[1,20,*]

[1]Department of Biomedical Sciences, University of Padova, I-35121 Padova, Italy, [2]Institute of Biosciences and Medical Technology, University of Tampere, Finland, [3]Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, 46202 Indianapolis, IN, USA, [4]Department of Woman and Child Health, University of Padova, I-35128 Padova, Italy, [5]Conway Institute of Biomolecular & Biomedical Research, University College Dublin, Belfield, Dublin 4, Ireland, [6]Ireland UCD School of Medicine & Medical Science, University College Dublin, Belfield, Dublin 4, Ireland, [7]Centre for Multidisciplinary Research, Institute of Nuclear Sciences Vinca, University of Belgrade, 11001 Belgrade, Serbia, [8]MTA-ELTE Lendület Bioinformatics Research Group, Department of Biochemistry, Eötvös Loránd University, 1/c Pázmány Péter sétány, 1117 Budapest, Hungary, [9]Institute of Enzymology, Research Centre for Natural Sciences, Hungarian Academy of Sciences, PO Box 7,H-1518 Budapest, Hungary, [10]Department of Biochemistry and Biophysics and Science for Life Laboratory, Stockholm University, Box 1031, 17121 Solna, Sweden, [11]Centre de Recherche en Biologie cellulaire de Montpellier (CRBM), UMR 5237 CNRS, Université Montpellier 1919 Route de Mende, Cedex 5, Montpellier 34293, France, [12]Institut de Biologie Computationnelle (IBC), Montpellier 34095, France, [13]University ITMO, Institute of Bioengineering, St. Petersburg 197101, Russia, [14]Department of Veterinary Medicine, University of Cambridge, Madingley Road, Cambridge CB3 0ES, UK, [15]Structural Biology Brussels, Vrije Universiteit Brussel (VUB), Brussels 1050, Belgium, [16]Structural Biology Research Center (SBRC), Flanders Institute for Biotechnology (VIB), Brussels 1050, Belgium, [17]Biomolecular Structure and Function Group, Instituto de Biologia Molecular e Celular (IBMC) and Instituto de Investigação e Inovação em Saúde (i3S), Universidade do Porto, 4200-135 Porto, Portugal, [18]Departament de Bioquimica i Biologia Molecular and Institut de Biotecnologia i Biomedicina, Universitat Autònoma de Barcelona, Bellaterra 08193, Spain, [19]Aix-Marseille Univ, CNRS, AFMB, UMR 7257, Marseille, France, [20]CNR Institute of Neurosceince, I-35121 Padova, Italy, [21]Interuniversity Institute of Bioinformatics in Brussels (IB2), ULB-VUB, Brussels 1050, Belgium, [22]Laboratory of Structural Dynamics, Stability and Folding of Proteins, Institute of Cytology, Russian Academy of Sciences, 194064 St. Petersburg, Russia and [23]Department of Molecular Medicine and USF Health Byrd Alzheimer's Research Institute, Morsani College of Medicine, University of South Florida, Tampa, FL 33612, USA

Received September 27, 2016; Revised October 19, 2016; Editorial Decision October 20, 2016; Accepted October 21, 2016

*To whom correspondence should be addressed. Tel: +39 049 827 6269; Email: silvio.tosatto@unipd.it
Correspondence may also be addressed to Sonia Longhi. Tel: +33 4 91 82 55 80; Email: Sonia.Longhi@afmb.univ-mrs.fr
Correspondence may also be addressed to Peter Tompa. Tel: +32 2 629 1962; Email: ptompa@vub.ac.be
†These authors contributed equally to the paper as first authors.

ABSTRACT

The Database of Protein Disorder (DisProt, URL: www.disprot.org) has been significantly updated and upgraded since its last major renewal in 2007. The current release holds information on more than 800 entries of IDPs/IDRs, i.e. intrinsically disordered proteins or regions that exist and function without a well-defined three-dimensional structure. We have re-curated previous entries to purge DisProt from conflicting cases, and also upgraded the functional classification scheme to reflect continuous advance in the field in the past 10 years or so. We define IDPs as proteins that are disordered along their entire sequence, i.e. entirely lack structural elements, and IDRs as regions that are at least five consecutive residues without well-defined structure. We base our assessment of disorder strictly on experimental evidence, such as X-ray crystallography and nuclear magnetic resonance (primary techniques) and a broad range of other experimental approaches (secondary techniques). Confident and ambiguous annotations are highlighted separately. DisProt 7.0 presents classified knowledge regarding the experimental characterization and functional annotations of IDPs/IDRs, and is intended to provide an invaluable resource for the research community for a better understanding structural disorder and for developing better computational tools for studying disordered proteins.

INTRODUCTION

Our traditional view of protein structure and function is deeply rooted in the structure–function paradigm which stated that the polypeptide chain of proteins needs to fold into a stable three-dimensional (3D) structure, which is a prerequisite of the functioning of the protein. The extreme explanatory power and success of this model is attested by more than hundred thousand high-resolution structures in the Protein Data Bank (PDB) (1) and many Nobel Prizes awarded for describing structures central to understanding important cell-biological phenomena. It has been suggested almost 20 years ago, however, that many proteins or regions of proteins in various proteomes lack such stable 3D structure, and are rather intrinsically disordered under native, physiological-like conditions (thus named IDPs/IDRs, respectively) (2–4). The recognition of this structural phenomenon brought a radical change in the structure–function paradigm, and critically extended the general appreciation of the role of dynamics in protein function. It has been recognized that structural disorder, which is prevalent in all organisms, plays roles primarily in cellular signaling and regulation (5). Because of that, IDPs/IDRs are often implicated in diseases (6) and represent important drug targets (7).

The structural and functional characterization of disordered proteins represents a special challenge, because they exist as an ensemble of rapidly interconverting conformations. Although they cannot be crystallized and thus cannot be directly characterized by X-ray crystallography, there are a variety of techniques that can report on their highly dynamic structural state at low- or even high spatial and temporal resolution (3). The current best structural description of IDPs/IDRs is by structural ensembles, which can be solved by a combination of experimental and computational approaches and are collected into a dedicated structural database, PED (8).

Studies of the structure–function relationship of disordered proteins have shown that in certain cases their function arises directly from the disordered state (entropic chains), whereas in many other cases their function emanates from molecular recognition accompanied by induced folding to specific binding partners, such as another protein, RNA or DNA molecule (9,10). In these functions, the sensitivity to regulated remodeling of the disordered structural ensemble is an excellent substrate for protein regulation, as exemplified by frequent post-translational modifications (11) and special modes of allosteric regulation (12) involving IDPs/IDRs.

Due to the prevalence and importance of structural disorder, several dedicated databases covering various aspects of IDPs/IDRs have appeared in the past decade. DisProt is the primary repository of disorder-related data on sequence- and functional annotations, focusing on disordered proteins or regions with experimental verification (13,14). Several other databases are based on predictions of disorder, such as D^2P^2, which contains disorder protein predictions by a variety of predictors on 1765 complete proteomes (15), MobiDB, which features three levels of annotations, manually curated, indirect and predicted for all UniProt sequences (over 80 million) (16), and IDEAL, which contains manual annotations of interaction regions undergoing induced folding, sites of post-translational modifications and assignments of structural domains (17). In addition, as already mentioned, PED is the database that gathers structural information on IDPs/IDRs, in the form of structural ensembles (8). The interaction of IDPs/IDRs with their target(s) is most often mediated by short continuous stretches of amino acids such as Molecular Recognition Elements/Features (MoREs/MoRFs) (18) and short/eukaryotic linear motifs (SLiMs/ELMs), which have been collected in the ELM database (19). Less frequently, partner interactions of IDPs/IDRs may also be mediated by intrinsically disordered domains (IDDs), i.e. longer regions that conform to the definition of domains as functional, evolutionary and structural units (20). Although probably still underappreciated, some of these IDDs may be found in the Pfam database of protein families which includes their annotations and underlying multiple sequence alignments (21).

DisProt is central to all IDP-related research efforts, because it collects and presents in a structured way the core experimental evidence reported for structural disorder in proteins. To give a new impetus to the field, we have significantly updated and upgraded it with new features. This new release—DisProt 7.0—contains more than 800 entries of IDPs/IDRs. We have also re-defined and extended functional categories laying the basis for a functional ontology of IDPs, now encompassing 7 major classes and 35 subclasses, all based on published experimental data.

Detection and characterization of IDPs

Technical advances in the field of biophysical and structural biology in the last 50 years have provided the scientific community with an arsenal of techniques to tackle the challenging characterization of IDPs/IDRs (4,22). The various methods differ in their extent of sophistication, and hence in their technical demand, as well as in the nature of the information they provide. Nuclear magnetic resonance (NMR) and X-ray crystallography provide site-specific information, whereas other methods provide more qualitative and global information (e.g. far-UV circular dichroism, size-exclusion chromatography; SEC).

The rise of the field of protein disorder has greatly benefited from structural biology, because structures deposited in the PDB (1) have been instrumental for the development of disorder predictors, often trained on regions of missing electron density. Developments of multidimensional heteronuclear NMR also enabled the structural characterization of disordered proteins of increasing size (23,24). In particular, heteronuclear single quantum coherence (HSQC) experiments are most commonly used to define protein disorder irrespective of whether residue-specific chemical shifts are available or not, as crowded HSQC spectra, characterized by a poor spread of resonances, are typical of IDPs/IDRs. The same feature of low spread of proton resonances is also apparent in one-dimensional proton-based NMR spectra, which offers the obvious advantage of not requiring isotopic labeling. Following assignment of the spectrum, quantitative estimations of disorder can be obtained through various NMR observables, such as chemical shifts, relaxation rates, residual dipolar couplings and resonance intensities in paramagnetic relaxation enhancement experiments. These data enable probing sequence-specific structural information in IDPs/IDRs. A particular strength of NMR is that it can be increasingly applied under truly *in vivo* conditions, in live cells (25). Therefore, these two experimental approaches, X-ray crystallography and multidimensional NMR, are considered as the 'primary techniques' providing evidence for structural disorder on a per residue basis in DisProt.

It should not miss our attention, though, that due to the expenses of isotopic labeling in NMR and the high rate of failure in protein crystallization, it would be unreasonable to only rely on these two approaches to document protein disorder. Therefore, beyond X-ray crystallography and NMR, a plethora of alternative biochemical and biophysical approaches (termed 'secondary techniques') provide orthogonal information on protein disorder in DisProt (4,22). The various approaches are of course not equivalent in terms of reliability, resolution and accuracy and suffer from specific drawbacks and limitations. Structural disorder is often based on far-UV CD spectroscopy, which is overall quite reliable, but does not enable discrimination between ordered and molten globular forms. Near-UV CD, beyond being able to unveil the lack of ordered structure, has the advantage of distinguishing between globular and molten globule forms. Another hallmark of disorder is anomalous sodium dodecyl sulphate-polyacrylamide gel electrophoresis migration, where IDPs have a high apparent molecular mass. IDPs/IDRs also behave anomalously in SEC, light scattering (DLS, MALS), and in small-angle X-ray scattering in that they display hydrodynamic radii (RH) and radii of gyration (Rg) higher than expected, reflecting an extended conformation.

Fluorescence spectroscopy is another common method to assess disorder. Intrinsic fluorescence probing the chemical environment of tryptophan residues provides information about their solvent-accessibility, whereas thermal differential scanning fluorimetry—similar to differential scanning calorimetry—can highlight the lack of a cooperative thermal transition and hence absence of ordered structure. Fluorescence resonance energy transfer between external fluorophores can even generate information on distance distributions and help solve the structural ensemble of the IDP (26). Hyper-sensitivity to proteolysis is also commonly used to map out disordered regions of proteins. Recently, native mass spectrometry exploiting nano-electrospray ionization (27,28) and high-speed atomic force microscopy operating at the single-molecule level (29) have emerged as attractive alternatives to address structural disorder.

As a last statement, it is noteworthy that the higher the number of independent experimental lines supporting disorder, the higher the reliability of the annotation. Furthermore, multi-dimensional information may help realize that structural disorder is not a single homogeneous structural state along an order-disorder binary classification coordinate, it rather represents a continuum of states from the fully ordered to the fully disordered. Similarly, many examples of biological relevant disorder in fragments that are missing from the full length protein have been reported. Furthermore, numerous functional examples of 'conditional disorder', i.e. instances where a disordered region functions by transitions to or from a folded state (30), or when disorder is only observed in a fraction of similar structures (31), lead to ambiguity and clearly points to the need for carrying out complementary experiments. In addition, an extreme case leading to conflicting results is represented by instances where a protein region, predicted to be ordered, is not defined in the electron density in one crystal structure while being ordered in another one (for an example see (32) and DisProt entry DP00133). Do these ambiguous regions represent a new class of disorder that escape detection using the currently available disorder predictors (thus setting the scene for their improvement), or *a contrario* are they the result of static disorder that arises from experimental conditions or domain wobbling? Combining information from a variety of sources may help clarify these cases and also improve meaningful descriptions of IDPs as conformational ensembles (33,34), which may lead to future descriptions of the structure–function relationship of IDPs.

Database structure and implementation

Database records. The technology of DisProt has been updated and is now based on a document-oriented MongoDB database. Stored documents are of two types, 'protein' including general information about the protein and 'disordered region (DR)' including evidence of disorder from literature. Protein information is retrieved from UniProt and includes cleavage sites and chain/peptide boundaries for polyproteins and processed proteins. DisProt is sequence-

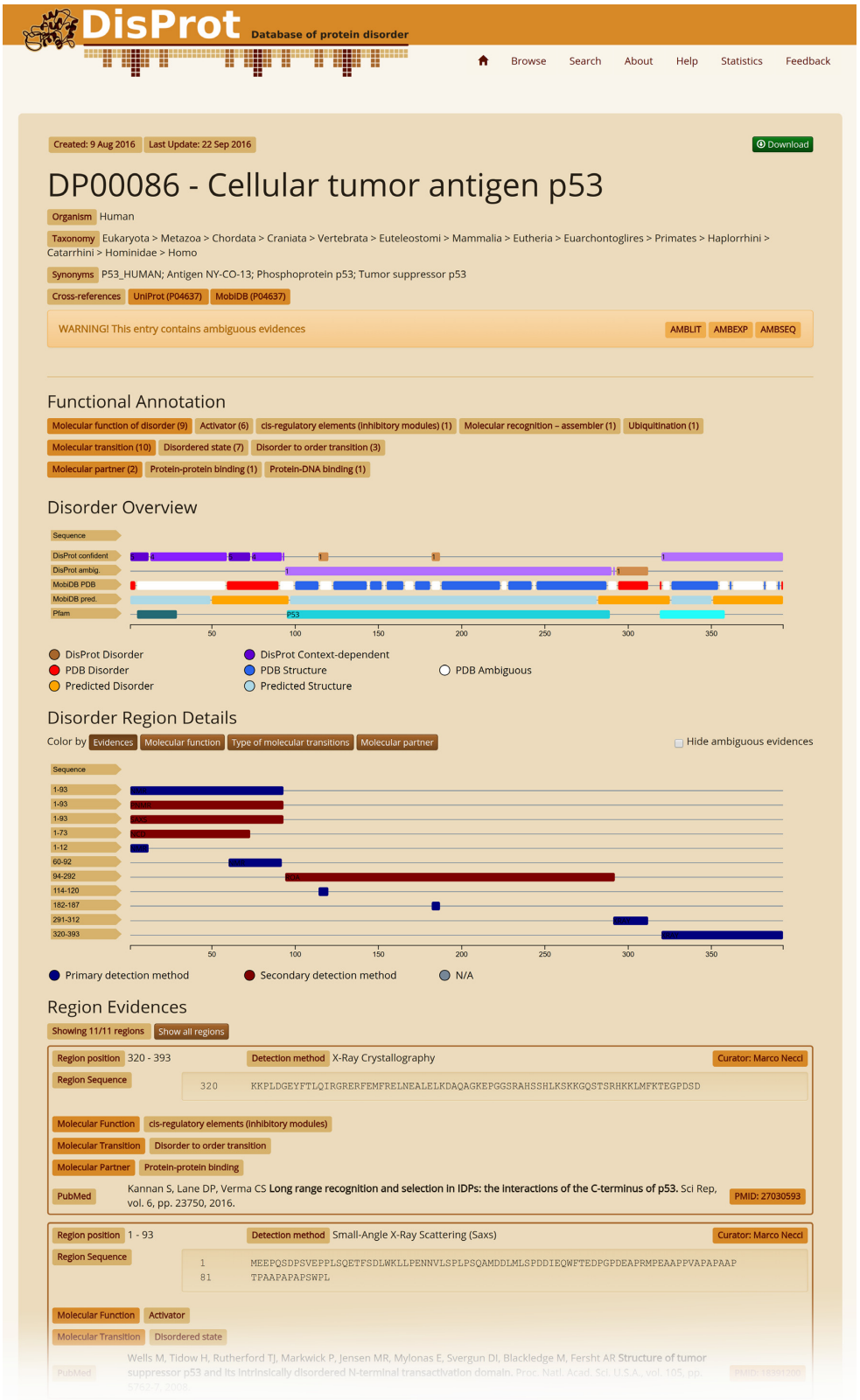

Figure 1. DisProt sample entry, human p53 protein (DP00086). Several experiments have been carried out to characterize the human p53 protein. DisProt reports literature evidence for IDRs. In particular, 11 different IDR evidences (Region Evidences) have been collected from nine different papers by two different curators. Most of these are related to the N-terminus and come from different types of experiments (Disorder Region Details). Disorder regions and the number of DisProt evidences, separated into confident and ambiguous annotations, can be compared with structural information from the Pfam and MobiDB databases in the Disorder Overview. DisProt also provides function annotation of IDRs by reporting molecular function, transition and partner terms (Functional Annotation). A literature reference is provided for each annotated IDR, linked to the relevant PubMed entry.

centric and different isoforms correspond to different entries as in the previous version. Cleaved proteins are merged into a single entry as they are products of the same native sequence. DisProt accession numbers now follow a single format and all previous entries with a '_xxx' suffix were removed. DR records are evidence-centric, i.e. different documents are stored for different experiments even when related to the same region. Forcing a one-to-one paradigm allows to track annotation evidence type and the corresponding literature source unambiguously. DR records also include experimental evidence quality tags for ambiguous annotations. Sometimes experiments are carried out on engineered sequences or fragments which may prove ambiguous to generalize for the entire sequence (AMBSEQ). Moreover, disorder boundaries are occasionally not clear from the literature (AMBLIT) or experiments are performed under extremely non-physiological conditions (AMBEXP). The major improvement from previous versions is the manually curated functional annotation of the regions. Whenever possible, curator-associated functions based on literature evidence are indicated by selecting terms from a new ontology built for describing disorder-related functional modes. If none of the current terms in the new ontology give a proper description of the functional mode, the curator may propose a new term to be added to the ontology. Acceptance of the new term will require approval by the IDP/IDR ontology committee.

Annotation pipeline. The new DisProt data have been generated by a community effort through a web server interface accessible upon registration. The same infrastructure can be used both to create and update entries. Curators provide an annotation through a submission form where all fields are validated on the client-side and a sequence viewer allows the comparison of assigned regions with structure information (Pfam domains, MobiDB disorder). Of note, the name of the curator is clearly visible in the entry to allow proper attribution of credit. The pipeline is fully automatic and can be potentially applied to the entire UniProt database. The DisProt public database is a snapshot of the community annotations.

Entry page. The entry page features four different sections (Figure 1). A protein information table gives the protein name, gene, synonyms, identifiers, taxonomy and 'homologous' entries inferred from sequence similarity. An interactive feature viewer reports DisProt disorder regions separated into confident and ambiguous annotations, colored brown for intrinsically disordered regions and purple for context-dependent regions. Pfam domains along with PDB and predicted disorder derived from MobiDB are also shown. Below, a detailed feature viewer provides different visualization layers to highlight different functional aspects (ontology terms) and the strength of available disorder evidence. Each position in the sequence is colored according to the number and type of evidence. Last but not least, the full curator-generated list of region evidences is reported on the bottom of the page and can be filtered by selecting an element (region) in the feature viewer. Figure 1 shows the current DisProt annotation for the human p53 protein. The combination of DisProt and PDB annotation clearly shows how p53 contains several segments undergoing disorder to order transitions. Evidence for disorder from the literature in the central p53 DNA binding domain, for which many crystal structures are available in the PDB, is ambiguous and highlighted with AMBLIT. Similar conflicts can probably be found in scores of DisProt entries and demonstrate the importance of flagging ambiguous data.

Browsing and searching data. Both browsing and searching functionalities are provided in a single solution from the 'Browse' page. A sortable, customizable and filterable table lists all entries by protein. Alternatively, another table listing all regions is available and accessible through the 'regions' button. Complex queries can be simulated applying different filters to different columns. Specific entries can be selected manually and customized views can be generated by adding or removing columns. Filtered and/or selected data can be downloaded both in text and JSON formats. Alternatively, the 'Search' page allows the user to search for specific words in a free-text form or to search for DisProt entries similar to a query sequence. Output for either search is a provided in a simplified form.

Feedback page. DisProt users are highly encouraged to suggest additional disorder annotations or changes to existing annotations using the 'Feedback' page. This contains a drop-down menu guiding the choice of feedback provided (e.g. website experience, novel annotations) and a message field. For feedback related to data entries, the user is asked to provide either the UniProt or DisProt ID and (where possible) a PubMed reference. All messages are reviewed by the curators and integrated in the database as time permits.

Web technology. The DisProt server is implemented in Node.js (https://nodejs.org) using the REST (Representational State Transfer) architecture. The data can be accessed through the web interface or programmatically exploiting the RESTful functionality. Please refer to the 'Help' section of the website for details on using the DisProt web services. The web interface is built using Angular.js (https://angularjs.org) and Bootstrap (http://getbootstrap.com) frameworks. The feature viewer is implemented on top of the Bio.js library.

Database content: upgrades and updates

Entries in DisProt 7.0 came from three major sources: (i) from the previous version of DisProt (where conflicting cases have been re-annotated), (ii) novel cases identified as PDB entries with long regions of missing electron density and (iii) proteins identified by text-mining in PubMed abstracts for keywords 'intrinsically disordered', 'intrinsically unstructured' and 'structural disorder'. New proteins selected based on disorder content (estimated based on MobiDB data) were prioritized (if appropriate information was available in SwissProt) to concentrate on well-studied and most interesting cases. New proteins were also selected by curators themselves to exploit their specific previous knowledge. All entries from previous versions were re-annotated to remove inconsistencies. One hundred and ninety-eight previous entries were completely removed and 469 modified. Recurring problems being fixed were wrong organism

Table 1. DisProt annotation content

Method/function	Proteins	Regions	Residues
Nuclear magnetic resonance (NMR)	333	592	32 926
X-ray crystallography	326	683	20 742
Circular dichroism (CD) spectroscopy, far-UV	261	352	53 935
Sensitivity to proteolysis	75	95	13 961
Size exclusion/gel filtration chromatography	62	67	12 206
Proton-based NMR	53	69	7723
SDS-PAGE gel, aberrant mobility on	34	34	6326
Other methods	237	273	41 833
Disorder transition	564	1505	151 498
Molecular function	489	1199	106 670
Molecular partner	444	1108	119 665

Distribution of DisProt annotation based on experimental evidence (method) and disorder function (function). As each annotated disorder region corresponds to one piece of experimental evidence, multiple regions can map to the same sequence segment. If a protein is annotated multiple times with the same type of experiment it is counted once. The number of residues is the sum of region lengths.

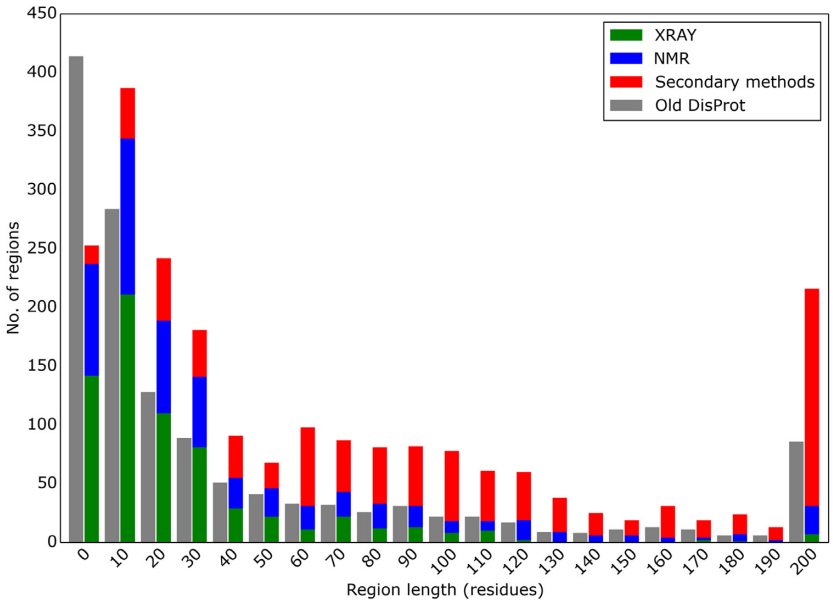

Figure 2. Distribution of disorder segment lengths. Segment lengths are binned in groups of 10 residues, e.g. the column 10 showing lengths between 10 and 19 residues. The current DisProt release is distinguished by experimental technique (X-ray in green, NMR in blue and other methods in red). The previous DisProt release is shown in a single gray bar as it did not have the experimental technique in a machine-readable format.

or isoform assignments, wrong IDR positioning, untracked disorder evidence (e.g. missing explicit literature reference) and weak evidence (e.g. based on very short fragments, please note that the minimal length of an IDR in DisProt 7.0 is 5 residues). Moreover, disorder annotations based on not traceable author/curator statements were discarded. Where necessary, a curator comment now highlights criticisms relative to a given evidence/experiment, e.g. if the experiment has been carried out on an engineered protein. Regions annotated as structured in previous DisProt releases were removed (33 regions). Information related to experiments has been simplified by skipping technical details regarding experimental conditions. However, weak experimental evidence is filtered out by the curator during annotation and tagged with one of three ambiguous labels. Overall, DisProt 7.0 includes 804 entries and 2167 disordered regions, with a total of 92 432 amino acids with clear experimental and functional annotations (Table 1), and the length distribution of disordered regions has significantly changed from the last release of DisProt (Figure 2).

New feature: functional classification

IDPs/IDRs carry out important functions in the cell. The field has settled on the notion that structural disorder represents a continuum of states from fully folded to fully unfolded (random coil-like), and function may come from any of the states and transitions between them. That is, their function may come directly from the disordered state or from molecular recognition and binding to partner molecule(s). We derive our classification from the logic of the gene ontology classification scheme (35), which is based on three structured ontologies ascribing functional terms to gene products (proteins) in terms of their associated biological processes (BP), cellular components (CC) and molecular functions (MF). Apparently, the CC and BP ontologies do not depend on the disordered status of the protein, they

Table 2. Major functional categories of the MFUN ontology of DisProt

MFUN code	Generic functional category	Functional category
MFUN_01	Entropic chain	Flexible linker/spacer
		Entropic bristle
		Entropic clock
		Entropic spring
		Structural mortar
		Self-transport through channel
MFUN_02	Molecular recognition: assembler	Assembler
		Localization (targeting)
		Localization (tethering)
		Prion (self-assembly, polymerization)
		Liquid-liquid phase separation/demixing (self-assembly)
MFUN_03	Molecular recognition: scavenger	Neutralization of toxic molecules
		Metal binding/metal sponge
		Water storage
MFUN_04	Molecular recognition: effector	Inhibitor
		Disassembler
		Activator
		cis-regulatory elements (inhibitory modules)
		DNA bending
		DNA unwinding
MFUN_05	Molecular recognition: display site	Phosphorylation
		Acetylation
		Methylation
		Glycosylation
		Ubiquitination
		Fatty acylation (myristolation and palmitoylation)
		Limited proteolysis
MFUN_06	Molecular recognition: chaperone	Protein detergent/solvate layer
		Space filling
		Entropic exclusion
		Entropy transfer

The functional schemes are an open hierarchy. One goal of sharing information with the community through DisProt is to refine our views of the functional modes of IDPs.

simply reflect the intracellular location of the protein and the BP it participates in, which can be kept without reference for the disordered status (35). The situation is entirely different with MF, which describes the elemental activities of a protein at the molecular level. In this regard, IDPs basically differ from folded proteins, such as enzymes or ligand-binding receptors, because their mode of action and type of function are usually completely different from those of folded proteins. Therefore, we have developed a novel classification scheme that merges and expands previous schemes that suggested thirty (36) and six (9) different categories, to provide classified descriptors for their MFs. Because previous categories (9,36) lacked coherence (for example, they treated structural transitions and interaction partners at the same level), we created a rational scheme that distinguishes these different types of ontologies (cf. Table 2 and ref. (3)).

The three sub-ontologies are as follows: (i) molecular function of disorder (MFUN): describes the type of functional readout of function (such as molecular chaperone); (ii) molecular transition (TRAN) necessary for function (such as disorder-to-order transition); and (iii) molecular partner (PART) that is recognized by the disordered protein (such as protein/RNA/DNA/small molecule). The MFUN ontology is described in detail in Table 1. The TRAN ontology can be further simplified to two IDR states (disorder and transition) to highlight different types of behavior, e.g. in the feature viewer of each DisProt entry.

CONCLUSIONS AND FUTURE WORK

We have presented an updated and completely re-worked version of the DisProt database. It now features state-of-the-art database and web technology, enabling programmatic access of interested parties. The content was expanded by defining a standardized set of experimental techniques and a novel functional ontology of disordered segments. Both allow for a richer description of disorder which may be used for further analyses. The other main improvement in DisProt is a complete re-annotation of existing entries to remove inconsistencies and an expansion of ca. 50% over the previous release, which also resulted in a significant shift in the length coverage of disordered regions in the database. This advance was made possible by a distributed annotation effort coordinated by the COST Action NGP-net (URL: ngp-net.bio.unipd.it) involving a dozen different groups and close to 40 annotators. The longer term maintenance of DisProt is provided by the Italian node of the European bioinformatics infrastructure Elixir. In the future we hope that DisProt can be able to provide disorder annotations for UniProt.

Finally, we hope that the upgrade of DisProt will encourage the scientific community to deposit experimental evidence for disorder within this unique repository, and that this renewed momentum will lead to an increased awareness of the importance of intrinsic disorder in proteins.

FUNDING

COST Action BM1405 NGP-net; ELIXIR-IIB (elixir-italy.org); 'Lendület' Grant from the Hungarian Academy of Sciences [LP2014-16 to Z.D.]; Hungarian Scientific Research Fund [OTKA K 108798 to Z.D.]; AIRC Research Fellowship (to D.P.); Spanish Ministerio de Educación Cultura i Deporte PhD Fellowship (to J.P.); Mexican National Council for Science and Technology (CONACYT) PhD Fellowship [215503 to M.M.-C.]; Grant PortoNeuroDRIve@i3S funded by Norte Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, through the European Regional Development Fund (ERDF) (to S.M.-R.); Direction Générale des Armées and Aix-Marseille University PhD Fellowship (to E.Sa.); OTKA Grant [PD-OTKA 108772 to E.Sc.]; French Ministry of National Education, Research and Technology PhD Fellowship (to A.S.); Ministry of Education, Science and Technological Development of the Republic of Serbia [173001, 173049 to N.V., R.D.]; ICREA-Academia Award (to S.V.); Odysseus Grant from Research Foundation Flanders (FWO) [G.0029.12 to P.T.]; AIRC IG Grant [17753 to S.T., in part]; Italian Ministry of Health [GR-2011-02347754 to E.L., S.T.; GR-2011-02346845 to S.T.]; Swedish Research Council Grant [VR-NT 2012-5046 to A.E.]. Computational resources were provided by the Swedish National Infrastructure for Computing (SNIC) at NSC. Funding for open access charge: COST Action BM1405 NGP-net.

Conflict of interest statement. None declared.

REFERENCES

1. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.
2. Oldfield,C.J. and Dunker,A.K. (2014) Intrinsically disordered proteins and intrinsically disordered protein regions. *Annu. Rev. Biochem.*, **83**, 553–584.
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. Habchi,J., Tompa,P., Longhi,S. and Uversky,V.N. (2014) Introducing protein intrinsic disorder. *Chem. Rev.*, **114**, 6561–6588.
5. Xie,H., Vucetic,S., Iakoucheva,L.M., Oldfield,C.J., Dunker,A.K., Uversky,V.N. and Obradovic,Z. (2007) Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions. *J. Proteome Res.*, **6**, 1882–1898.
6. Uversky,V.N., Oldfield,C.J. and Dunker,A.K. (2008) Intrinsically disordered proteins in human diseases: introducing the D2 concept. *Annu. Rev. Biophys.*, **37**, 215–246.
7. Metallo,S.J. (2010) Intrinsically disordered proteins are potential drug targets. *Curr. Opin. Chem. Biol.*, **14**, 481–488.
8. Varadi,M., Kosol,S., Lebrun,P., Valentini,E., Blackledge,M., Dunker,A.K., Felli,I.C., Forman-Kay,J.D., Kriwacki,R.W., Pierattelli,R. *et al.* (2014) pE-DB: a database of structural ensembles of intrinsically disordered and of unfolded proteins. *Nucleic Acids Res.*, **42**, D326–D335.
9. Tompa,P. (2005) The interplay between structure and function in intrinsically unstructured proteins. *FEBS Lett.*, **579**, 3346–3354.
10. Wright,P.E. and Dyson,H.J. (2009) Linking folding and binding. *Curr. Opin. Struct. Biol.*, **19**, 31–38.
11. Iakoucheva,L.M., Radivojac,P., Brown,C.J., O'Connor,T.R., Sikes,J.G., Obradovic,Z. and Dunker,A.K. (2004) The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res.*, **32**, 1037–1049.
12. Tompa,P. (2014) Multisteric regulation by structural disorder in modular signaling proteins: an extension of the concept of allostery. *Chem. Rev.*, **114**, 6715–6732.
13. Vucetic,S., Obradovic,Z., Vacic,V., Radivojac,P., Peng,K., Iakoucheva,L.M., Cortese,M.S., Lawson,J.D., Brown,C.J., Sikes,J.G. *et al.* (2005) DisProt: a database of protein disorder. *Bioinformatics*, **21**, 137–140.
14. Sickmeier,M., Hamilton,J.A., LeGall,T., Vacic,V., Cortese,M.S., Tantos,A., Szabo,B., Tompa,P., Chen,J., Uversky,V.N. *et al.* (2007) DisProt: the Database of Disordered Proteins. *Nucleic Acids Res.*, **35**, D786–D793.
15. 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.
16. Potenza,E., Di Domenico,T., Walsh,I. and Tosatto,S.C. (2015) MobiDB 2.0: an improved database of intrinsically disordered and mobile proteins. *Nucleic Acids Res.*, **43**, D315–D320.
17. Fukuchi,S., Sakamoto,S., Nobe,Y., Murakami,S.D., Amemiya,T., Hosoda,K., Koike,R., Hiroaki,H. and Ota,M. (2012) IDEAL: intrinsically disordered proteins with extensive annotations and literature. *Nucleic Acids Res.*, **40**, D507–D511.
18. Mohan,A., Oldfield,C.J., Radivojac,P., Vacic,V., Cortese,M.S., Dunker,A.K. and Uversky,V.N. (2006) Analysis of molecular recognition features (MoRFs). *J. Mol. Biol.*, **362**, 1043–1059.
19. Dinkel,H., Van Roey,K., Michael,S., Kumar,M., Uyar,B., Altenberg,B., Milchevskaya,V., Schneider,M., Kuhn,H., Behrendt,A. *et al.* (2016) ELM 2016–data update and new functionality of the eukaryotic linear motif resource. *Nucleic Acids Res.*, **44**, D294–D300.
20. Tompa,P., Fuxreiter,M., Oldfield,C.J., Simon,I., Dunker,A.K. and Uversky,V.N. (2009) Close encounters of the third kind: disordered domains and the interactions of proteins. *Bioessays*, **31**, 328–335.
21. Finn,R.D., Coggill,P., Eberhardt,R.Y., Eddy,S.R., Mistry,J., Mitchell,A.L., Potter,S.C., Punta,M., Qureshi,M., Sangrador-Vegas,A. *et al.* (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.*, **44**, D279–D285.
22. Receveur-Brechot,V., Bourhis,J.M., Uversky,V.N., Canard,B. and Longhi,S. (2006) Assessing protein disorder and induced folding. *Proteins*, **62**, 24–45.
23. Kosol,S., Contreras-Martos,S., Cedeno,C. and Tompa,P. (2013) Structural characterization of intrinsically disordered proteins by NMR spectroscopy. *Molecules*, **18**, 10802–10828.
24. Felli,I.C. and Pierattelli,R. (2012) Recent progress in NMR spectroscopy: toward the study of intrinsically disordered proteins of increasing size and complexity. *IUBMB Life*, **64**, 473–481.
25. Theillet,F.X., Binolfi,A., Bekei,B., Martorana,A., Rose,H.M., Stuiver,M., Verzini,S., Lorenz,D., van Rossum,M., Goldfarb,D. *et al.* (2016) Structural disorder of monomeric alpha-synuclein persists in mammalian cells. *Nature*, **530**, 45–50.
26. Schuler,B., Soranno,A., Hofmann,H. and Nettels,D. (2016) Single-molecule FRET spectroscopy and the polymer physics of unfolded and intrinsically disordered Proteins. *Annu. Rev. Biophys.*, **45**, 207–231.
27. Kaltashov,I.A., Bobst,C.E. and Abzalimov,R.R. (2013) Mass spectrometry-based methods to study protein architecture and dynamics. *Protein Sci.*, **22**, 530–544.
28. Borysik,A.J., Kovacs,D., Guharoy,M. and Tompa,P. (2015) Ensemble methods enable a new definition for the solution to gas-phase transfer of intrinsically disordered proteins. *J. Am. Chem. Soc.*, **137**, 13807–13817.
29. Miyagi,A., Tsunaka,Y., Uchihashi,T., Mayanagi,K., Hirose,S., Morikawa,K. and Ando,T. (2008) Visualization of intrinsically disordered regions of proteins by high-speed atomic force microscopy. *Chemphyschem*, **9**, 1859–1866.
30. Jakob,U., Kriwacki,R. and Uversky,V.N. (2014) Conditionally and transiently disordered proteins: awakening cryptic disorder to regulate protein function. *Chem. Rev.*, **114**, 6779–6805.
31. DeForte,S. and Uversky,V.N. (2016) Resolving the ambiguity: making sense of intrinsic disorder when PDB structures disagree. *Protein Sci.*, **25**, 676–688.
32. Blocquel,D., Habchi,J., Durand,E., Sevajol,M., Ferron,F., Erales,J., Papageorgiou,N. and Longhi,S. (2014) Coiled-coil deformations in crystal structures: the measles virus phosphoprotein multimerization

domain as an illustrative example. *Acta Crystallogr. D Biol. Crystallogr.*, **70**, 1589–1603.
33. Sterckx,Y.G., Volkov,A.N., Vranken,W.F., Kragelj,J., Jensen,M.R., Buts,L., Garcia-Pino,A., Jove,T., Van Melderen,L., Blackledge,M. *et al.* (2014) Small-angle X-ray scattering- and nuclear magnetic resonance-derived conformational ensemble of the highly flexible antitoxin PaaA2. *Structure*, **22**, 854–865.
34. Aznauryan,M., Delgado,L., Soranno,A., Nettels,D., Huang,J.R., Labhardt,A.M., Grzesiek,S. and Schuler,B. (2016) Comprehensive structural and dynamical view of an unfolded protein from the combination of single-molecule FRET, NMR, and SAXS. *Proc. Natl. Acad. Sci. U.S.A.*, **113**, E5389–E5398.
35. Consortium, G.O. (2015) Gene ontology consortium: going forward. *Nucleic Acids Res.*, **43**, D1049–D1056.
36. Dunker,A.K., Brown,C.J., Lawson,J.D., Iakoucheva,L.M. and Obradovic,Z. (2002) Intrinsic disorder and protein function. *Biochemistry*, **41**, 6573–6582.