PED in 2021: a major update of the protein ensemble database for intrinsically disordered proteins

Tamas Lazar, Elizabeth Martínez-Pérez, Federica Quaglia, András Hatos, Lucía Chemes, Javier Iserte, Nicolás Méndez, Nicolás Garrone, Tadeo Saldaño, Julia Marchetti, et al.

D404–D411 *Nucleic Acids Research, 2021, Vol. 49, Database issue* *Published online 10 December 2020*
https://doi.org/10.1093/nar/gkaa1021

PED in 2021: a major update of the protein ensemble database for intrinsically disordered proteins

Tamas Lazar [1,2], Elizabeth Martínez-Pérez [3,4], Federica Quaglia [5], András Hatos [5], Lucía B. Chemes [6], Javier A. Iserte [3], Nicolás A. Méndez [6], Nicolás A. Garrone [6], Tadeo E. Saldaño [7], Julia Marchetti [7], Ana Julia Velez Rueda [7], Pau Bernadó [8], Martin Blackledge [9], Tiago N. Cordeiro [8,10], Eric Fagerberg [11], Julie D. Forman-Kay [12,13], Maria S. Fornasari [7], Toby J. Gibson [4], Gregory-Neal W. Gomes [14,15], Claudiu C. Gradinaru [14,15], Teresa Head-Gordon [16], Malene Ringkjøbing Jensen [9], Edward A. Lemke [17,18], Sonia Longhi [19], Cristina Marino-Buslje [3], Giovanni Minervini [5], Tanja Mittag [20], Alexander Miguel Monzon [5], Rohit V. Pappu [21], Gustavo Parisi [7], Sylvie Ricard-Blum [22], Kiersten M. Ruff[21], Edoardo Salladini [19], Marie Skepö [11,23], Dmitri Svergun [24], Sylvain D. Vallet [22], Mihaly Varadi [25], Peter Tompa [1,2,26,*], Silvio C.E. Tosatto [5,*] and Damiano Piovesan [5]

[1]VIB-VUB Center for Structural Biology, Flanders Institute for Biotechnology, Brussels 1050, Belgium, [2]Structural Biology Brussels, Bioengineering Sciences Department, Vrije Universiteit Brussel, Brussels 1050, Belgium, [3]Bioinformatics Unit, Fundación Instituto Leloir, Buenos Aires, C1405BWE, Argentina, [4]Structural and Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg 69117, Germany, [5]Dept. of Biomedical Sciences, University of Padua, Padova 35131, Italy, [6]Instituto de Investigaciones Biotecnológicas "Dr. Rodolfo A. Ugalde", IIB-UNSAM, IIBIO-CONICET, Universidad Nacional de San Martín, CP1650 San Martín, Buenos Aires, Argentina, [7]Laboratorio de Química y Biología Computacional, Departamento de Ciencia y Tecnología, Universidad Nacional de Quilmes, Bernal B1876BXD, Buenos Aires, Argentina, [8]Centre de Biochimie Structurale (CBS), CNRS, INSERM, University of Montpellier, Montpellier 34090, France, [9]Univ. Grenoble Alpes, CNRS, CEA, IBS, Grenoble, F-38000, France, [10]Instituto de Tecnologia Química e Biológica António Xavier, Universidade Nova de Lisboa, Av. da República, Oeiras 2780-157, Portugal, [11]Theoretical Chemistry, Lund University, Lund, POB 124, SE-221 00, Sweden, [12]Molecular Medicine Program, Hospital for Sick Children, Toronto, M5G 1X8, Ontario, Canada, [13]Department of Biochemistry, University of Toronto, Toronto, M5S 1A8, Ontario, Canada, [14]Department of Physics, University of Toronto, Toronto, M5S 1A7, Ontario, Canada, [15]Department of Chemical and Physical Sciences, University of Toronto Mississauga, Mississauga, L5L 1C6, Ontario, Canada, [16]Departments of Chemistry, Bioengineering, Chemical and Biomolecular Engineering University of California, Berkeley, CA 94720, USA, [17]Biocentre, Johannes Gutenberg-University Mainz, Mainz 55128, Germany, [18]Institute of Molecular Biology, Mainz 55128, Germany, [19]Aix-Marseille University, CNRS, Architecture et Fonction des Macromolécules Biologiques (AFMB), Marseille 13288, France, [20]Department of Structural Biology, St. Jude Children's Research Hospital, Memphis, TN 38105, USA, [21]Department of Biomedical Engineering, Center for Science & Engineering of Living Systems (CSELS), Washington University in St. Louis, MO 63130, USA, [22]Univ Lyon, University Claude Bernard Lyon 1, CNRS, INSA Lyon, CPE, Institute of Molecular and Supramolecular Chemistry and Biochemistry (ICBMS), UMR 5246, Villeurbanne, 69629 Lyon Cedex 07, France, [23]LINXS - Lund Institute of Advanced Neutron and X-ray Science, Lund 223 70, Sweden, [24]European Molecular Biology Laboratory, Hamburg Unit, Hamburg 22607, Germany, [25]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, CB10 1SD, UK and [26]Institute of Enzymology, Research Centre for Natural Sciences, Budapest, 1117, Hungary

Received September 14, 2020; Revised October 13, 2020; Editorial Decision October 14, 2020; Accepted December 08, 2020

*To whom correspondence should be addressed. Tel +32 473 785386; Email: peter.tompa@vub.be
Correspondence may also be addressed to Silvio C. E. Tosatto. Tel: +39 049 827 6269; Email: silvio.tosatto@unipd.it

ABSTRACT

The Protein Ensemble Database (PED) (https://proteinensemble.org), which holds structural ensembles of intrinsically disordered proteins (IDPs), has been significantly updated and upgraded since its last release in 2016. The new version, PED 4.0, has been completely redesigned and reimplemented with cutting-edge technology and now holds about six times more data (162 versus 24 entries and 242 versus 60 structural ensembles) and a broader representation of state of the art ensemble generation methods than the previous version. The database has a completely renewed graphical interface with an interactive feature viewer for region-based annotations, and provides a series of descriptors of the qualitative and quantitative properties of the ensembles. High quality of the data is guaranteed by a new submission process, which combines both automatic and manual evaluation steps. A team of biocurators integrate structured metadata describing the ensemble generation methodology, experimental constraints and conditions. A new search engine allows the user to build advanced queries and search all entry fields including cross-references to IDP-related resources such as DisProt, MobiDB, BMRB and SASBDB. We expect that the renewed PED will be useful for researchers interested in the atomic-level understanding of IDP function, and promote the rational, structure-based design of IDP-targeting drugs.

INTRODUCTION

Valuable mechanistic and functional information can be obtained from protein structures modeled at atomistic resolution (1–3). Due to the growth of experimentally determined structures deposited in the Protein Data Bank (PDB) (4), currently there are >160 000 3D structures of macromolecules available in the database (4). As structural biology has mainly focused on determining the structure of globular proteins until the recent past, the presence of intrinsically disordered protein (IDP) regions (IDRs) have mostly been inferred either from unresolved or proteolytically digested tails or loops of these globular structures solved by X-ray crystallography, or from shorter regions yielding few structural constraints in nuclear magnetic resonance (NMR) spectroscopy measurements (5). The depletion of long IDRs (LDRs) in PDB has been known for a long time, and the tightening of the gap in this regard has only become practical very recently (6). However, this recent abundance of LDRs is predominantly due to the context-dependent folding of proteins with conditional disorder, such as pH sensitivity, PTM-dependent folding, localization-dependent disorder and folding upon binding to a partner (7–9).

Although conditionally folded IDRs provide important structural insights, in-depth understanding of mechanistic details of how IDPs function also requires knowledge about the dynamic structures in the free state. By virtue of their extreme conformational dynamics, ensemble description is often applied for structural modeling of IDPs. Conformational ensembles are representative sets of conformers reflecting on the structural dynamics of IDPs sampling the space. Ensemble modeling usually relies on experimental data originating from NMR spectroscopy (10–13) and small-angle X-ray scattering (SAXS) data (14–18), Förster resonance energy transfer (FRET) (19,20) circular dichroism (CD) spectroscopy data (21) or a combination thereof (22–25). These measurements are then used to define local or nonlocal structural constraints for the computational modeling of the conformational ensemble, such as for the restraining or reweighting of a pool of statistical random coils, or of molecular dynamics (MD) trajectories (22,26–28).

Solving structural ensembles, however, is fraught with uncertainties, because the number of degrees of freedom is inherently much larger than the number of experimentally determined structural restraints. As a result, determining an ensemble is a mathematically 'ill-posed' or 'underdetermined' problem that has more than one solution. We don't yet know how to select the 'best' ensemble from multiple alternatives, neither can we be sure if an actual ensemble is a faithful representation of the real physical state of the IDP/IDR, nor is only a reasonable fit to experiment observations. To help address these issues, IDP/IDR ensembles solved at the time were collected and made available in the dedicated Protein Ensemble database (published as pE-DB in 2014 (26), renamed as PED in later versions).

This first version was an ambitious attempt to fill the niche in the deposition of ensembles of fully disordered proteins and proteins with IDRs. At the time of the publication, it only stored data for a few dozens of ensembles for a limited set of proteins, which increased very slowly in the following years. Manual deposition and validation of entry submissions used to hinder the smooth maintenance and increment of the database. A lot has happened, however, in the structural–functional characterization of IDPs/IDRs since the inception of PED. For example, it has been proven that structural ensembles can predict independent structural data (24,26), i.e. they are realistic and do have predictive power. Based on novel, better suited force-fields (29–32) the computational simulation of IDPs has also significantly advanced (33,34). Influential IDP-related databases have been either updated (e.g. DisProt (35) or MobiDB (36)) or created anew (e.g. MFIB (37) or DIBS (38)). Successful targeting of IDPs/IDRs by small molecules offers hope for a new class of effective drugs (39,40). A superposition-free method for comparing alternative ensembles has also been worked out (41), and allosteric regulatory mechanisms operating in the heterogeneous ensemble of IDPs/IDRs (multistery) have been elaborated (42,43). The appreciation of the importance of structural disorder in the novel field of liquid–liquid phase separation (LLPS) is on the rise (44) and persistent structural disorder of phase-separating proteins even in the condensed state has been reported (45,46). Last, but not least, many ensembles have been solved (24,47) but not made publicly available.

This rapid progress in the protein disorder field mandates a basic upgrade and significant update of PED. To meet this goal, PED 4.0 was completely redesigned and extended with

several new functionalities. To set a higher standard for the quality of data, a new submission process is now carried out through a web interface that enables automated validation of the ensemble deposited by the authors and manual curation steps with the assistance of the database biocurators. PED is now better cross-referenced with other IDP-related databases such as BMRB (48), SASBDB (49), DisProt (35) and MobiDB (36), and has a well-documented RESTful API for programmatic access, search and download. In all, the new PED has about six times more data than the previous version.

PROGRESS

Database structure and implementation

One of the major changes since the previous version is the whole new deposition process, which includes an automatic data validation step and a curation step. The validation has been introduced to standardize the data and improve its quality by providing a number of structural indicators, while the manual curation step provides metadata for better data accessibility. A team of biocurators standardize the description of the experimental methodology using terms from a controlled vocabulary and identify cross-references to third-party databases. Curators also scan the literature to collect ensembles that have not yet been deposited into PED.

In PED, an entry is identified by the PED prefix and 5 digits (e.g. PED00001), which corresponds to an experiment on a protein (or protein complex), while a PED ensemble (e.g. PED00001e001) is the set of conformations (or models) generated to fit the experimental data. Different ensembles generated using the same proteoform (same sequence construct and PTMs), the same experimental conditions and the same experimental and computational methodology, represent different replicas of the same experiment (alternative solutions to the same set of structural restraints determined) and are grouped together in the same PED entry.

PED also stores conformation weights as provided by the authors. Weights represent the probability for each conformation to populate the ensemble, however, since these are not yet standardized and provided only for a limited number of entries, they have not been considered in the calculation of ensemble descriptors such as Rg, accessibility and secondary structure propensities.

The backend of the PED server processes each entry submission. The server executes a collection of scripts developed in-house that generate summary statistics (solvent accessibility, secondary structure populations, radius of gyration and maximum dimension). Secondary structure and solvent accessibility are calculated by DSSP (50,51), while MolProbity (52) provides quality descriptors (torsion-angle outliers, covalent bond-length and angle outliers, beta-carbon deviations and steric clashes). For each entry, the pipeline generates a report, which can be used to assess a submission. Since the same approach is used for all entries, it is possible to make comparisons across the entire database and generate meaningful statistics. The report is available for download as a PDF document for all entries.

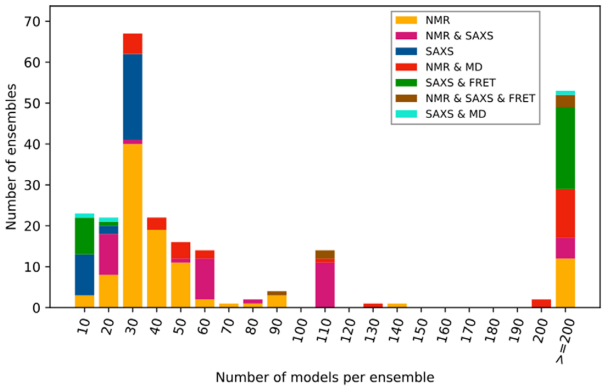

Figure 1. PED 4.0 entry statistics. Stacked histogram of models per ensemble for different measurement methods in PED 4.0, binned based on the number of the consisting conformer models.

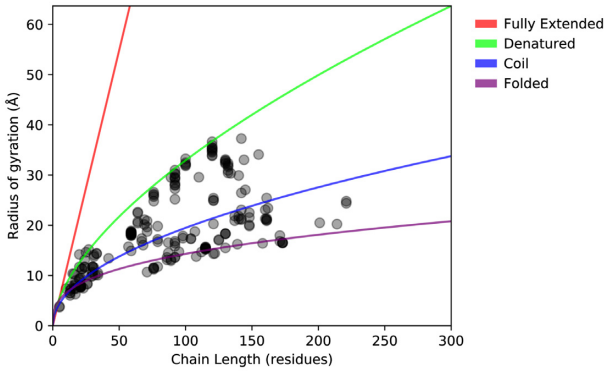

Figure 2. Chain compactness of PED 4.0 entries. Radius of gyration of protein chains plotted against their chain length. Each dot represents a given chain in a given ensemble. The reference curves (54) represent values specific for folded proteins (purple), random coils (blue), denatured proteins (green) and fully extended chains (red). Four long folded proteins (PED00007, PED00010, PED00014 and PED00162) with over 300 residues are omitted, but fit well to the purple trend line.

DATABASE CONTENT

New entries

The number of entries in PED 4.0 has increased six-fold compared to the previous release. Some entries have been deposited after literature curation, while others have been directly provided by the experimentalists who generated the data (data owners). Previous entries were manually reviewed and re-annotated. Old entries that included different experiments were split up. The mapping from old to new identifiers is reported on the website (URL: https://proteinensemble.org/help#mapping).

For new entries, PED curators focused on biologically interesting protein regions with conformational ensembles, or more often, a set of ensembles determined under different conditions (different construct or mutant, different pH, denaturants etc.) or using different types of experimental datasets and modeling methodology. As sensitivity to conditions is well-known for IDPs, these alternate ensembles might provide very valuable insights into the conditional disorder of these proteins (41). Furthermore, multiple ensembles for a region measured under very similar condi-

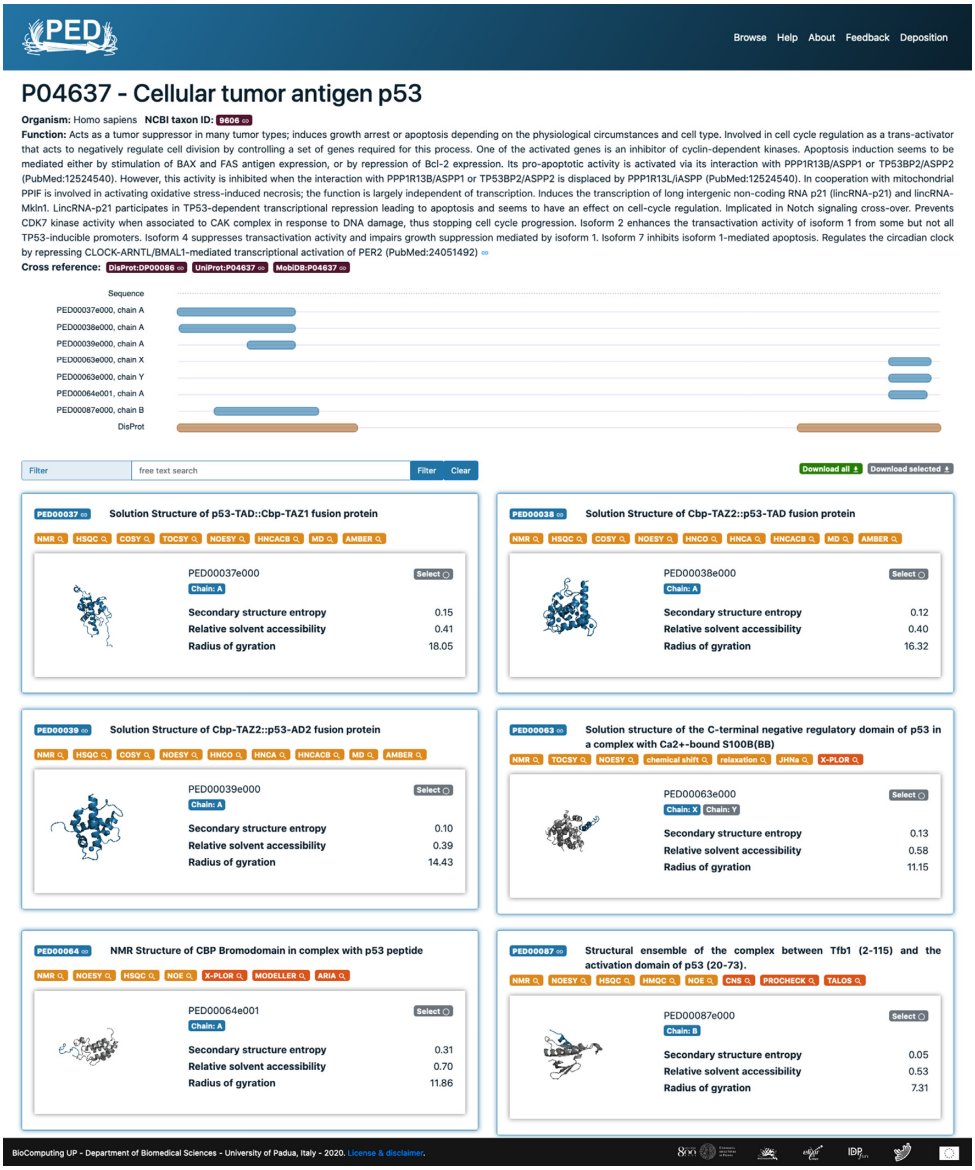

Figure 3. Example for PED's Protein page. Protein page P04637 summarizes the human p53 ensembles currently stored in PED for both the N-terminal and C-terminal disordered region. The feature viewer also integrates intrinsic disorder evidence from DisProt.

tions may highlight the biases in the modeling protocols and procedures (41). The curation efforts focused primarily on the submission of larger conformational ensembles (min. 40 conformers for a given protein region) of preferentially IDRs determined using experimental constraints. The DisProt database was harnessed to make sure many of the additions correspond to bona fide experimentally determined IDPs/IDRs. This was complemented by an analysis of the radii of gyration (R_g) of the protein chains.

Statistics

Statistical analysis of the PED entries (Figure 1) shows an increment in all classes of determination methods and in all sizes of ensembles (i.e. number of models ranging from a dozen to thousands). It also highlights that while NMR remains as the most highly represented method used to model ensembles of usually <100 conformers, other state-of-the-art methods such as SAXS and smFRET are also represented in the new dataset. Large ensembles (number of conformers >50) determined by SAXS only are absent in PED, as currently most SAXS-based ensemble modeling tools are known to represent ensembles through several equivalent data sets, reducing the number of representative models in each set to a range of 10–50 (15,53). However, larger ensembles generated by a combination of SAXS, smFRET and molecular dynamics methods are present.

Protein compactness is often characterized by the R_g as a function of the length of the polypeptide chain (Figure 2). The R_g of folded proteins scales with chain length by following a scaling law while, trivially, rigid rod-like chains follow a linear trend. Disordered proteins, however, fall in between these two extremes due to their propensities to form local or nonlocal transient secondary (or tertiary) structure ele-

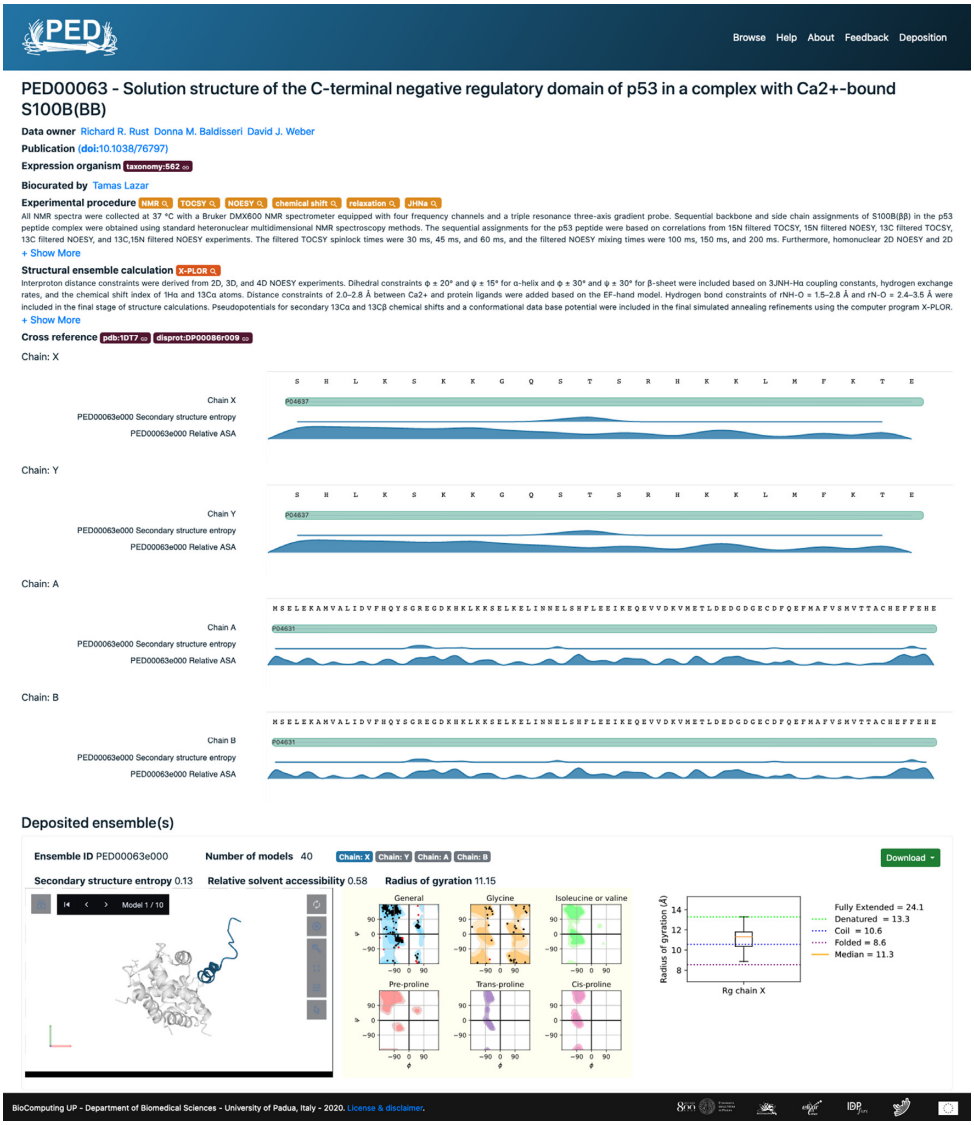

Figure 4. Example for PED's Entry page. Entry page is shown for the C-terminal disordered region of p53 in a tetrameric complex with Ca^{2+}-bound S100B (PED00063). The feature viewer shows chain-specific information, while molecular graphics, Ramachandran maps and R_g distribution are presented below.

ments. Figure 2 shows that the disordered proteins of PED largely exhibit an R_g ranging from that of random coils to that of denatured proteins (54) across a wide range of IDP protein lengths (10–200 residues), implying that the IDPs in PED represent the known variety of IDP compaction behavior. The points lying on the folded line correspond to globular binding partners present in the ensembles that represent complexes of IDPs and folded proteins/domains.

NOVEL FEATURES

Now PED 4.0 has both a protein-centric and an experimental entry-centric view. In the protein-centric view (Protein page), ensembles from different PED entries are grouped based on their UniProt accession. In this way, it is possible to appreciate the differences between ensembles corresponding to the same region on a single page, which may arise from the use of different techniques and conditions. The 'Entry' page provides details about the experimental design and shows information on the complete make-up of the ensemble, i.e. describes if a protein complex includes nonpeptidic molecules or protein chains not mapping to UniProt (55).

Figure 3 shows the 'Protein' page for human p53 with multiple available ensembles for both the N-terminal and C-terminal regions. By clicking on PED identifiers, it is possible to open the corresponding Entry pages. For example, PED00063 (Figure 4) corresponds to the p53 C-terminal region folding upon binding to S100B. In sharp contrast, PED00064 (not shown) is a disordered complex of p53 binding to the CBP bromodomain.

Entry views

The Entry page (Figure 4) provides the title of the experiment, authors, and the corresponding publication when

available. PED does not include primary data, like structural constraints, but instead provides cross-references to primary databases; when available (PDB (4), BMRB (48) and SASBDB (49)). MobiDB (36) and DisProt (35) are cross-referenced in order to link evidence about the intrinsic disorder of the protein region.

For each entry, the PED biocurators generate a detailed description of the ensemble determination. This description about experimental and computational protocols is organized into three different blocks (experimental procedure, structural ensemble calculation and, if applicable, MD calculations), each including a narrative and a set of terms selected from a controlled vocabulary (CV). The CV ensures advanced accessibility and searchability and is constantly updated to capture new developments of the field. The current CV is available on the 'About' page of the PED website.

The rest of the Entry page provides a graphical view of structural features of the ensemble. The Feature-Viewer (56) component summarizes the make-up at the chain level. It shows the protein construct, solvent accessibility, secondary-structure populations and the respective variability (entropy or standard deviation) across ensemble models. For each chain of the ensemble, the distribution of the radii of gyration (R_g) is shown as a box plot, along with the corresponding theoretical values (dashed lines) for a protein chain of the same length if it was folded, random coil-like or denatured, and expected R_g value for a rod-like or fully extended chain of the same length. Torsion angles are mapped to a Ramachandran plot to evaluate the structural preferences of the ensemble of the entire protein complex (not chains) and the quality of backbone modeling. A Quick view on the ensemble conformations (models) is provided by the MOL* structure viewer (57). The metadata, ensemble coordinates and validation report are all downloadable.

Browse and search

Browse and advance search are implemented on the same page. A customizable table lists all entries with information about the protein, types of measurements, number of ensembles and conformers. Each row represents a chain of an ensemble or a fragment in cases when the ensemble is calculated on an engineered construct. The corresponding UniProt accessions are provided for the majority of the PED entries. A search box allows the user to look up specific words in a free-text form or to search PED and all cross-referenced identifiers. Moreover, it is possible to search all the terms from the controlled vocabulary and to build complex queries or exploit regular expressions. Simple search is also available on the Main page, while programmatic search and data access (or download) is implemented via a RESTful API. An extended documentation and examples are provided on the Help page.

CONCLUSION

After several years of steadily diminishing activity, PED has finally come to new life. First, it has been transferred to a stable location that ensures continuous maintenance and regular updates, hopefully stimulating the development of novel approaches—experimental and computational tools—for developing and depositing ever more accurate ensembles. Second, it has been significantly extended in size and has a greatly improved representation of ensemble-generation methodologies and of functionally validated 'bona fide' IDRs, thanks to a community-wide curation effort. The number of entries has increased from 24 to 152, whereas the number of ensembles has grown from 60 to 215. In all, the total number of 'conformers' stored in the database now exceed 290,532 PDB models (versus 24 615 in the old PED).

PED has also been profoundly upgraded in a quest for better consistency. The most important novel feature is the implementation of a new deposition process aimed at improving the quality of the entire database. PED now includes a web submission system. Each deposition is subjected to an automatic validation step, which generates a report on model quality, and a manual curation step, in which a submission is manually evaluated and integrated with structured metadata. The automatic validation step includes statistics on bond angles and lengths, backbone torsion angles and steric clashes. Whereas statistics on 'outliers' in the various geometric categories do not entail the rejection of deposition, it gives the user the option of selecting only ensembles that meet certain preset quality criteria. The biocurator submission interface will soon be made available to the public with the idea of providing a tool similar to the OneDep system of the wwPDB (4) in the near future. Contributing new ensembles is highly encouraged, and for that, information about submission inquiries are available on the Deposition page.

Additional novel features of PED 4.0 include a completely new implementation of the website and database schema. PED stores ensemble weights representing conformational probabilities. Even though these are not taken into account in the calculation of ensemble properties due to a lack of standardization, they will be extensively integrated in the future. The web interface has both a protein- and experiment-centric view, an advanced search engine and a well-documented API for programmatic access.

The quick and significant growth of PED is due to the steady activity of experimentalists generating disordered ensembles that accumulated large amounts of data in the past years, and the perseverance of database curators. This signals the vitality of the concept of protein disorder and the strength of the disorder community working on integrating structural, functional and medical aspects of structural disorder. We expect that the new database will foster a significant conceptual leap in the field. Even today, after more than two decades of research that has brought solidification of the basic concept, we still tend to perceive structural disorder as a binary classifier, thinking of proteins or protein regions as either ordered or disordered. Structural disorder, however, is not a simple, homogeneous structural state, it rather represents a continuum of states from fully ordered to fully disordered (58). PED is currently the only database focused on representing the diversity of IDP protein ensembles, which are not stored in databases focused on the deposition of primary data (SASDB, BMRB, PDB), creating an extremely valuable resource for the IDP community. The analysis of ensembles in PED 4.0 will enable us to better un-

derstand determinants of these various sub-states in terms of compactness, secondary structure content and dynamics, which will definitely help correctly interpret the functional consequences of intrinsic structural disorder. Given the prevalence of structural disorder in disease (59), the insight expected from structural ensembles in PED 4.0 will also give a new impetus to efforts of structure-based drug discovery against IDPs.

The renewal and relocation of the database reflects on the ambition of the IDP community to actively maintain the database and, more ambitiously, also to integrate it into DisProt's IDP-specific complex ecosystem of databases and computational tools (60). Significant further developments in the near future, such as mirroring the database among multiple locations and contacting journals to recommend ensemble deposition into PED, are also planned. Continuous maintenance and implementation of these and other future plans are ensured by the IDPcentral, MSCA-RISE IDPfun and ELIXIR IDP community groups. To ensure communication with users about recent growth of the database and new features, PED now will have a more active social media presence on Twitter with the original @ProteinEnsemble Twitter account.

ACKNOWLEDGEMENTS

PED is maintained as a service of the ELIXIR IDP community (URL: elixir-europe.org/communities/intrinsically-disordered-proteins). This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 778247.

FUNDING

Italian Ministry of University and Research (MIUR) to SCET [2017483NH8]; European Union's Horizon 2020 to SCET [778247]; Hungarian Scientific Research Fund (OTKA) to PT [K124670, K131702]; Universidad Nacional de Quilmes to GP [PUNQ 1309/19]; National Agency for the Promotion of Science and Technology (ANPCyT) to GP [PICT-2014-3430] and to LBC [PICT-2017-1924]; Fondation pour la Recherche Médicale to SRB and SDV [DBI20141231336]; Natural Sciences and Engineering Research Council of Canada to CCG [RGPIN 2017-06030]; Agence Nationale de la Recherche (ANR) to PB [ANR-10-LABX-12-01]; National Institutes of Health (NIH) to JFK and THG [5R01GM127627-03]; German Ministry of Science and Education (SAS-BSOFT) to DS [16QK10A]; EU Horizon 2020 programme (iNEXT-Discovery) to DS [871037]; Vrije Universiteit Brussel (VUB) to PT [SRP51]; EM-P, NG, NM, JM are PhD students, AJVR, TES are Postdocs and GP, CM-B, JI, LBC and MSF are researchers of the National Research Council (CONICET) of Argentina. Funding for open access charge: Vrije Universiteit Brussel (VUB) [SRP51].

Conflict of interest statement. None declared.

REFERENCES

1. PDBe-KB,consortium. (2020) PDBe-KB: a community-driven resource for structural and functional annotations. *Nucleic Acids Res.*, **48**, D344–D353.
2. Sillitoe,I., Dawson,N., Lewis,T.E., Das,S., Lees,J.G., Ashford,P., Tolulope,A., Scholes,H.M., Senatorov,I., Bujan,A. *et al.* (2019) CATH: expanding the horizons of structure-based functional annotations for genome sequences. *Nucleic Acids Res.*, **47**, D280–D284.
3. Mitchell,A.L., Attwood,T.K., Babbitt,P.C., Blum,M., Bork,P., Bridge,A., Brown,S.D., Chang,H.-Y., El-Gebali,S., Fraser,M.I. *et al.* (2019) InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res.*, **47**, D351–D360.
4. wwPDB,consortium. (2019) Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res.*, **47**, D520–D528.
5. Tompa,P. (2010) In: *Structure and function of intrinsically disordered proteins*. Chapman & Hall/CRC Press, Boca Raton.
6. Monzon,A.M., Necci,M., Quaglia,F., Walsh,I., Zanotti,G., Piovesan,D. and Tosatto,S.C.E. (2020) Experimentally determined long intrinsically disordered protein regions are now abundant in the Protein Data Bank. *Int. J. Mol. Sci.*, **21**, 143–143.
7. Bugge,K., Brakti,I., Fernandes,C.B., Dreier,J.E., Lundsgaard,J.E., Olsen,J.G., Skriver,K. and Kragelund,B.B. (2020) Interactions by Disorder - A matter of context. *Front. Mol. Biosci.*, **7**, 110.
8. Hausrath,A.C. and Kingston,R.L. (2017) Conditionally disordered proteins: bringing the environment back into the fold. *Cell. Mol. Life Sci.*, **74**, 3149–3162.
9. 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.
10. Ozenne,V., Schneider,R., Yao,M., Huang,J., Salmon,L., Zweckstetter,M., Jensen,M.R. and Blackledge,M. (2012) Mapping the potential energy landscape of intrinsically disordered proteins at amino acid resolution. *J. Am. Chem. Soc.*, **134**, 15138–15148.
11. Kosol,S., Contreras-Martos,S., Cedeño,C. and Tompa,P. (2013) Structural characterization of intrinsically disordered proteins by NMR spectroscopy. *Mol. Basel Switz.*, **18**, 10802–10828.
12. Jensen,M.R., Zweckstetter,M., Huang,J. and Blackledge,M. (2014) Exploring free-energy landscapes of intrinsically disordered proteins at atomic resolution using NMR spectroscopy. *Chem. Rev.*, **114**, 6632–6660.
13. Salvi,N., Salmon,L. and Blackledge,M. (2017) Dynamic descriptions of highly flexible molecules from NMR dipolar Couplings: Physical basis and limitations. *J. Am. Chem. Soc.*, **139**, 5011–5014.
14. Cordeiro,T.N., Herranz-Trillo,F., Urbanek,A., Estaña,A., Cortés,J., Sibille,N. and Bernadó,P. (2017) Small-angle scattering studies of intrinsically disordered proteins and their complexes. *Curr. Opin. Struct. Biol.*, **42**, 15–23.
15. Tria,G., Mertens,H.D.T., Kachala,M. and Svergun,D.I. (2015) Advanced ensemble modelling of flexible macromolecules using X-ray solution scattering. *IUCrJ*, **2**, 207–217.
16. Gräwert,T.W. and Svergun,D.I. (2020) Structural modeling using solution Small-Angle X-ray Scattering (SAXS). *J. Mol. Biol.*, **432**, 3078–3092.
17. Vallet,S.D., Miele,A.E., Uciechowska-Kaczmarzyk,U., Liwo,A., Duclos,B., Samsonov,S.A. and Ricard-Blum,S. (2018) Insights into the structure and dynamics of lysyl oxidase propeptide, a flexible protein with numerous partners. *Sci. Rep.*, **8**, 11768.
18. Hamdi,K., Salladini,E., O'Brien,D.P., Brier,S., Chenal,A., Yacoubi,I. and Longhi,S. (2017) Structural disorder and induced folding within two cereal, ABA stress and ripening (ASR) proteins. *Sci. Rep.*, **7**, 15544.
19. Holmstrom,E.D., Holla,A., Zheng,W., Nettels,D., Best,R.B. and Schuler,B. (2018) Accurate transfer efficiencies, distance distributions, and ensembles of unfolded and intrinsically disordered proteins from Single-Molecule FRET. *Methods Enzymol.*, **611**, 287–325.
20. Fuertes,G., Banterle,N., Ruff,K.M., Chowdhury,A., Mercadante,D., Koehler,C., Kachala,M., Estrada Girona,G., Milles,S., Mishra,A. *et al.* (2017) Decoupling of size and shape fluctuations in heteropolymeric sequences reconciles discrepancies in SAXS vs. FRET measurements. *Proc. Natl. Acad. Sci. U.S.A.*, **114**, E6342–E6351.
21. Nagy,G., Igaev,M., Jones,N.C., Hoffmann,S.V. and Grubmüller,H. (2019) SESCA: Predicting circular dichroism spectra from protein molecular structures. *J. Chem. Theory Comput.*, **15**, 5087–5102.

22. Krzeminski,M., Marsh,J.A., Neale,C., Choy,W.-Y. and Forman-Kay,J.D. (2013) Characterization of disordered proteins with ENSEMBLE. *Bioinformatics*, **29**, 398–399.
23. Sterckx,Y.G.J., Volkov,A.N., Vranken,W.F., Kragelj,J., Jensen,M.R., Buts,L., Garcia-Pino,A., Jové,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. *Struct. Lond. Engl. 1993*, **22**, 854–865.
24. Schwalbe,M., Ozenne,V., Bibow,S., Jaremko,M., Jaremko,L., Gajda,M., Jensen,M.R., Biernat,J., Becker,S., Mandelkow,E. *et al.* (2014) Predictive atomic resolution descriptions of intrinsically disordered hTau40 and α-synuclein in solution from NMR and small angle scattering. *Struct. Lond. Engl. 1993*, **22**, 238–249.
25. Ibáñez de Opakua,A., Merino,N., Villate,M., Cordeiro,T.N., Ormaza,G., Sánchez-Carbayo,M., Diercks,T., Bernadó,P. and Blanco,F.J. (2017) The metastasis suppressor KISS1 is an intrinsically disordered protein slightly more extended than a random coil. *PLoS One*, **12**, e0172507.
26. 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–D35.
27. Rangan,R., Bonomi,M., Heller,G.T., Cesari,A., Bussi,G. and Vendruscolo,M. (2018) Determination of structural ensembles of Proteins: Restraining vs Reweighting. *J. Chem. Theory Comput.*, **14**, 6632–6641.
28. Köfinger,J., Stelzl,L.S., Reuter,K., Allande,C., Reichel,K. and Hummer,G. (2019) Efficient ensemble refinement by reweighting. *J. Chem. Theory Comput.*, **15**, 3390–3401.
29. Huang,J., Rauscher,S., Nawrocki,G., Ran,T., Feig,M., de Groot,B.L., Grubmüller,H. and MacKerell,A.D. (2017) CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods*, **14**, 71–73.
30. Song,D., Luo,R. and Chen,H.-F. (2017) The IDP-Specific force field ff14IDPSFF improves the conformer sampling of intrinsically disordered proteins. *J. Chem. Inf. Model.*, **57**, 1166–1178.
31. Robustelli,P., Piana,S. and Shaw,D.E. (2018) Developing a molecular dynamics force field for both folded and disordered protein states. *Proc. Natl. Acad. Sci. U.S.A.*, **115**, E4758–E4766.
32. Rahman,M.U., Rehman,A.U., Liu,H. and Chen,H.-F. (2020) Comparison and evaluation of force fields for intrinsically disordered proteins. *J. Chem. Inf. Model.*, **60**, 4912–4923.
33. Chong,S.-H., Chatterjee,P. and Ham,S. (2017) Computer simulations of intrinsically disordered proteins. *Annu. Rev. Phys. Chem.*, **68**, 117–134.
34. Shrestha,U.R., Juneja,P., Zhang,Q., Gurumoorthy,V., Borreguero,J.M., Urban,V., Cheng,X., Pingali,S.V., Smith,J.C., O'Neill,H.M. *et al.* (2019) Generation of the configurational ensemble of an intrinsically disordered protein from unbiased molecular dynamics simulation. *Proc. Natl. Acad. Sci. U.S.A.*, **116**, 20446–20452.
35. 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.
36. Piovesan,D., Tabaro,F., Paladin,L., Necci,M., Micetic,I., Camilloni,C., Davey,N., Dosztányi,Z., Mészáros,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.
37. Fichó,E., Reményi,I., Simon,I. and Mészáros,B. (2017) MFIB: a repository of protein complexes with mutual folding induced by binding. *Bioinforma. Oxf. Engl.*, **33**, 3682–3684.
38. Schad,E., Fichó,E., Pancsa,R., Simon,I., Dosztányi,Z. and Mészáros,B. (2018) DIBS: a repository of disordered binding sites mediating interactions with ordered proteins. *Bioinforma. Oxf. Engl.*, **34**, 535–537.
39. Tsafou,K., Tiwari,P.B., Forman-Kay,J.D., Metallo,S.J. and Toretsky,J.A. (2018) Targeting intrinsically disordered transcription Factors: Changing the paradigm. *J. Mol. Biol.*, **430**, 2321–2341.
40. 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.
41. Lazar,T., Guharoy,M., Vranken,W., Rauscher,S., Wodak,S.J. and Tompa,P. (2020) Distance-Based metrics for comparing conformational ensembles of intrinsically disordered proteins. *Biophys. J.*, **118**, 2952–2965.
42. Tompa,P. (2014) Multisteric regulation by structural disorder in modular signaling proteins: an extension of the concept of allostery. *Chem. Rev.*, **114**, 6715–6732.
43. Wodak,S.J., Paci,E., Dokholyan,N.V., Berezovsky,I.N., Horovitz,A., Li,J., Hilser,V.J., Bahar,I., Karanicolas,J., Stock,G. *et al.* (2019) Allostery in its many Disguises: From theory to applications. *Struct. Lond. Engl. 1993*, **27**, 566–578.
44. Brangwynne,C.P., Tompa,P. and Pappu,R.V. (2015) Polymer physics of intracellular phase transitions. *Nat. Phys.*, **11**, 899–904.
45. Murthy,A.C., Dignon,G.L., Kan,Y., Zerze,G.H., Parekh,S.H., Mittal,J. and Fawzi,N.L. (2019) Molecular interactions underlying liquid-liquid phase separation of the FUS low-complexity domain. *Nat. Struct. Mol. Biol.*, **26**, 637–648.
46. Murthy,A.C. and Fawzi,N.L. (2020) The (un)structural biology of biomolecular liquid-liquid phase separation using NMR spectroscopy. *J. Biol. Chem.*, **295**, 2375–2384.
47. Delaforge,E., Kragelj,J., Tengo,L., Palencia,A., Milles,S., Bouvignies,G., Salvi,N., Blackledge,M. and Jensen,M.R. (2018) Deciphering the dynamic interaction profile of an intrinsically disordered protein by NMR exchange spectroscopy. *J. Am. Chem. Soc.*, **140**, 1148–1158.
48. Romero,P.R., Kobayashi,N., Wedell,J.R., Baskaran,K., Iwata,T., Yokochi,M., Maziuk,D., Yao,H., Fujiwara,T., Kurusu,G. *et al.* (2020) BioMagResBank (BMRB) as a Resource for Structural Biology. *Methods Mol. Biol. Clifton NJ*, **2112**, 187–218.
49. Kikhney,A.G., Borges,C.R., Molodenskiy,D.S., Jeffries,C.M. and Svergun,D.I. (2020) SASBDB: Towards an automatically curated and validated repository for biological scattering data. *Protein Sci. Publ. Protein Soc.*, **29**, 66–75.
50. Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
51. Touw,W.G., Baakman,C., Black,J., te Beek,T.A.H., Krieger,E., Joosten,R.P. and Vriend,G. (2015) A series of PDB-related databanks for everyday needs. *Nucleic Acids Res.*, **43**, D364–D368.
52. Williams,C.J., Headd,J.J., Moriarty,N.W., Prisant,M.G., Videau,L.L., Deis,L.N., Verma,V., Keedy,D.A., Hintze,B.J., Chen,V.B. *et al.* (2018) MolProbity: More and better reference data for improved all-atom structure validation. *Protein Sci. Publ. Protein Soc.*, **27**, 293–315.
53. Bernadó,P., Mylonas,E., Petoukhov,M.V., Blackledge,M. and Svergun,D.I. (2007) Structural characterization of flexible proteins using small-angle X-ray scattering. *J. Am. Chem. Soc.*, **129**, 5656–5664.
54. Hofmann,H., Soranno,A., Borgia,A., Gast,K., Nettels,D. and Schuler,B. (2012) Polymer scaling laws of unfolded and intrinsically disordered proteins quantified with single-molecule spectroscopy. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 16155–16160.
55. UniProt Consortium (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.*, **47**, D506–D515.
56. Paladin,L., Schaeffer,M., Gaudet,P., Zahn-Zabal,M., Michel,P.-A., Piovesan,D., Tosatto,S.C.E. and Bairoch,A. (2020) The Feature-Viewer: a visualization tool for positional annotations on a sequence. *Bioinforma. Oxf. Engl.*, **36**, 3244–3245.
57. Sehnal,D., Rose,A.S., Koča,J., Burley,S.K. and Velankar,S. (2018) Mol*: towards a common library and tools for web molecular graphics. In: *Proceedings of the Workshop on Molecular Graphics and Visual Analysis of Molecular Data, MolVA '18*. Eurographics Association, Goslar, DEU, pp. 29–33.
58. Sormanni,P., Piovesan,D., Heller,G.T., Bonomi,M., Kukic,P., Camilloni,C., Fuxreiter,M., Dosztanyi,Z., Pappu,R.V., Babu,M.M. *et al.* (2017) Simultaneous quantification of protein order and disorder. *Nat. Chem. Biol.*, **13**, 339–342.
59. 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.
60. 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. *F1000Res.*, **8**, ELIXIR-1753.