Exploring Manually Curated Annotations of Intrinsically Disordered Proteins with DisProt

Exploring manually curated annotations of intrinsically disordered proteins with DisProt

Federica Quaglia[1], András Hatos[1] , Damiano Piovesan[1] and Silvio C.E. Tosatto[1, 2]
[1] Department of Biomedical Sciences, University of Padova, Padova 35121, Italy.
[2] Corresponding author: silvio.tosatto@unipd.it

ABSTRACT:

DisProt is the major repository of manually curated data for intrinsically disordered proteins collected from the literature. Although lacking a stable tertiary structure under physiological conditions, intrinsically disordered proteins carry out a plethora of biological functions, some of them directly arising from their flexible nature. A growing number of scientific studies has been published during the last few decades in an effort to shed light on their unstructured state, their binding modes and their functions. DisProt makes use of a team of expert biocurators to provide up-to-date annotations of intrinsically disordered proteins from the literature, making them available to the scientific community. Here we present a comprehensive description on how to use DisProt in different contexts and provide a detailed explanation on how to explore and interpret manually curated annotations of intrinsically disordered proteins. We describe how to search DisProt annotations, both using the web interface and the API for programmatic access. Finally, we explain how to visualize and interpret a DisProt entry, p53, which is a well-characterized intrinsically disordered protein.

Basic Protocol 1: Performing a search in DisProt
Support protocol 1: Downloading options
Support protocol 2: Programmatic access with DisProt REST API
Basic Protocol 2: Visualizing and interpreting DisProt entries - the p53 use case
Basic Protocol 3: Providing feedback and submitting new intrinsic disorder-related data
Guidelines for understanding results

KEYWORDS:

database - DisProt - literature curation - community curation - intrinsically disordered proteins

INTRODUCTION:

Intrinsically disordered proteins (IDPs) are characterised by the presence of unstructured and highly flexible segments, termed “intrinsically disordered regions” (IDRs), that lack a stable tertiary structure. IDRs can be easily detected by several biophysical and biochemical methods, among which X-ray and NMR are the most commonly used (Tompa, 2009; van der Lee et al., 2014). Missing electron density regions that can not be detected on X-ray crystal structures are due to unobserved atoms that fail to properly scatter X-rays, denoting their structural flexibility (Mészáros et al., 2007; Uversky & Dunker, 2010). NMR spectroscopy studies are also widely used to assess the presence of unstructured protein segments, being able to recognize disordered regions that in crystal structures are visible due to the formation of crystal contacts (Kobe et al., 2008; Mizutani et al., 2008). Several additional methods can assess the presence of intrinsic disorder in a protein, such as circular dichroism, sensitivity to proteolysis and small angle X-ray scattering (Uversky & Dunker, 2010). Intrinsically disordered proteins can also exist as partially structured folding intermediates, pre-molten globules and molten globules, that exhibit a higher degree of secondary structure than random coils while being less compact than native structures (Ptitsyn, 1995; Ptitsyn et al., 1995; van der Lee et al., 2014). They can play a crucial role in several biological processes, such as membrane localization and interaction with protein chaperones, to name a few (Uversky & Dunker, 2010). A main feature of IDPs is their peculiar mode of interaction. Interaction surfaces of IDPs are characterised by a unique set of chemico-physical properties, e.g. a higher percentage of hydrophobic residues compared to the rest of the IDR (Jones & Thornton, 1996; Mészáros et al., 2007). They exhibit a larger exposed interaction area per residue, even in their folded state induced by binding, that they use to contact their physiological partners (Mészáros et al., 2007). The lack of structure in IDR segments in their unbound state provides a multiplicity of advantages due to their largely extended conformation, such as 1) the possibility for a single IDR to be involved in interactions with more structurally different partners, 2) several structured partners being able to bind to a single region, 3) an increased speed of interaction due to their ability to explore the interaction space and 4) a reduced binding strength that allows for transient

interactions (Mészáros et al., 2007, 2009; Uversky & Dunker, 2010). IDRs can undergo a disorder-to-order transition upon binding of a partner, enabling them to play a central role as protein hubs (Mészáros et al., 2007, 2011), as in the case of p53 (DisProt identifier: DP00086) and α-synuclein (DisProt identifier: DP00070). Finally, IDPs can also be involved in the regulation of several biological processes, interacting with different types of binding partners such as proteins, nucleic acids, lipids and small molecules, therefore acting as molecular recognition effectors and assemblers (Cumberworth et al., 2013; Tompa, 2002, 2005; van der Lee et al., 2014). Strikingly, some of the most well characterized and crucial functions of IDPs arise from their flexible nature: they can be flexible linkers connecting structured domains of a protein, or they can act as entropic clocks, bristles and springs due to their entropic features (Tompa, 2009; Uversky & Dunker, 2010).

DisProt is a service of the Italian node of ELIXIR, the European infrastructure for biological data, and a key resource for the recently established ELIXIR IDP user community (Davey et al., 2019). It is also the largest repository of manually curated annotations of intrinsically disordered proteins (IDPs) collected from the literature (Hatos et al., 2020; Piovesan et al., 2017). A team of expert DisProt curators looks for new data on IDPs/IDRs from relevant publications and annotates them through a dedicated curation interface by means of intrinsic disorder-related annotation terms that are codified into the IDP ontology. The IDP ontology (URL: disprot.org/about) includes four main branches, corresponding to the four disorder aspects annotated in DisProt: structural state, structural transition, interaction partner and disorder function. A DisProt entry corresponds to a protein isoform and unambiguously maps to a UniProt entry. DisProt annotations describe local properties of the protein sequence (e.g. intrinsically disordered regions) which are always supported by experimental evidence taken from the literature. Each DisProt annotation is uniquely identified by the DisProt entry accession number followed by a suffix starting with a lowercase *r* letter (example DP00086r003).

In this article, we provide detailed protocols explaining how to perform a search in DisProt (Basic Protocol 1), visualize and interpret annotations of a DisProt entry (Basic Protocol 2), and how to submit a new evidence of intrinsic disorder in DisProt (Basic Protocol 3). We concurrently describe the downloading options in DisProt (Support Protocol 1) and programmatic access with the DisProt REST API (Support Protocol 2).

PERFORMING A SEARCH IN DISPROT

DisProt is freely accessible at URL: disprot.org/. This protocol describes how to search entries and to retrieve information in DisProt. From the home page users can also navigate the DisProt blog (URL: disprot.org/blog) to read posts describing our updates or explore the DisProt Twitter account (URL: twitter.com/disprot_db) (Fig. 1).

Necessary resources

Hardware

While DisProt works best on laptop or desktop computers, it is also easily accessible from smartphones and tablets. An active and stable internet connection is required.

Software

Internet browser, e.g. Firefox (URL: www.mozilla.org/firefox/), Google Chrome (URL: www.google.com/chrome/), Safari (URL: www.apple.com/safari/).

Input data

Free text search against the database.

Performing a text search

1. Open a web browser and connect to DisProt, URL: disprot.org/.
2. Searches in DisProt can be performed both using the “Search” boxes on the top-right and top-middle of the DisProt home page, or by clicking on the “Browse” button available on ther top-left of the home page.
    a. Users can perform a search using the “Search” boxes on the top-right or top-middle of the DisProt home page to look for protein entries or entries referencing a specific publication (Fig. 2). *Users can look for a specific protein, e.g. Alpha-synuclein from Homo sapiens, by submitting the protein name, e.g. Alpha-synuclein, or its corresponding UniProtKB accession number, P37840. They will be redirected to the corresponding DisProt entry, in this case DisProt identifier DP00070. Users could also be interested in looking for a specific publication. In this case, please enter in the search box*

the corresponding PubMed identifier (PMID) of the publication. All entries that have at least one evidence referencing that publication will be displayed.

b. Alternatively, it is possible to perform an advanced search by clicking on the "Browse" button available on the top-left of the home page (Fig. 3). Users will be redirected to an advanced search page, where they can refine their search and look for a specific query, or a combination of them, e.g. a protein name and an organism.

3. Select "Text search" on the top-left side of the Browse page, then select a term from the drop-down menu. *Users can look for the following aspects:*
   a. *A specific protein: please select a "Protein name", e.g. "Alpha-synuclein", and "UniProt ACC, e.g. P37840.*
   b. *A set of proteins from a specific organism: choose an "Organism", e.g. "Gallus gallus", the "Taxon" or "NCBI Taxon ID".*
   c. *UniProt Reference Clusters (UniRef). UniRef databases cluster UniProtKB sequences by gathering together proteins based on their sequence similarity (Suzek et al., 2015). Terms available are "UniRef50", "UniRef90" and "UniRef100" (clustering the sequences at 50%, 90% and 100% identity, respectively).*
   d. *Entries from a specific curator: select the "Curator name" term and start typing the name you are looking for.*
   e. *A specific reference: users can look for a specific PMID, e.g. 8632448, by selecting the "Reference ID" term or for the title of the corresponding publication, e.g. "Alternative arrangements of the protein chain are possible for the adenovirus single-stranded DNA binding protein", by selecting the "Reference name" term.*
   f. *A specific term from our ontology: select one among "Disorder Ontology ID" or "Disorder Ontology name". Users that wish to have a better insight on the terms of our ontology and read their descriptions, can refer to the Disorder Ontology description available from URL: disprot.org/about.*
   g. *It is also possible to perform a "free text" search by selecting the corresponding term in the drop-down menu.*
4. It is possible to customize the table columns to visualize more details of an entry in the displayed results. Default columns include "*DisProt ID*", "*UniProt ACC*", "*protein name*", "*organism*" and "*disorder content*". We suggest to add at least the "*annotated terms*" column to have an insight on the disorder aspects available for each entry.
5. Download the search results using the "Download selected" button at the top-left of the Browse page. File formats available for download are JSON, TSV or FASTA. Users can also choose to include ambiguous and/or obsolete entries by selecting the corresponding buttons above "Download selected".

Performing a sequence similarity search

1. Open a web browser and connect to DisProt: disprot.org/.
2. Click on the "Browse" button on the top-left side of the home page (Fig. 4) to be redirected to the advanced search page.
3. Select "BLAST" on the top-left side of the Browse page to perform a BLAST (Altschul et al., 1990) sequence similarity search against DisProt entries.
4. Insert a protein sequence in the corresponding box and click on "Submit".
5. DisProt entries that match the query will be displayed in the results. Available columns are "*DisProt ID*", "*UniProt ACC*", "*protein name*", "*organism*" and "*disorder content*" along with "*Bit-score*", "*E-value*", "*Identity*" and "*Coverage*".
6. Download the results of the search using the "Download selected" button at the top-left of the Browse page. File formats available for download are JSON, TSV or FASTA.

DOWNLOADING OPTIONS

Users can download a specific release of DisProt, e.g. 2016_10 (Disprot 7), or a specific version of the IDP ontology from the Download page (Fig. 5), URL: disprot.org/download.

Necessary resources

Hardware

While laptop or desktop computers are recommended, download options are also available when using a tablet or a smartphone. An active and stable internet connection is required.

Software

Internet browser, e.g. Firefox (URL: www.mozilla.org/firefox/), Google Chrome (URL: www.google.com/chrome/), Safari (URL: www.apple.com/safari/).

Input data

No input data are required.

Downloading a release of DisProt

1. Open a web browser and connect to DisProt: disprot.org/.
2. Click on the "Download" button on the top bar of DisProt.
3. Select the release of interest from the "Release" drop-down menu, e.g. "*2016_10 (Disprot 7)*".
4. Select a file format of the output file from the "Format" drop-down menu. Available options are: JSON, TSV and FASTA formats.
5. It is possible to include ambiguous and/or obsoleted regions by selecting them from the "Include" options, otherwise leave the corresponding boxes unchecked.
6. Click on the "Download" button.

Downloading the IDP ontology

1. Open a web browser and connect to DisProt: disprot.org/.
2. Click on the "Download" button on the top bar of DisProt.
3. Select a version of the ontology from the "Ontology" drop-down menu.
4. Select the file format of the output file. Available options are: JSON, OWL and OBO formats. OBO and OWL formats correspond to the Biomedical Ontology and Web Ontology Language, respectively.
5. Click on the "Download" button.

The downloadable output files described here are identical to those obtainable with the DisProt REST API, described in detail in the Support protocol 2.

PROGRAMMATIC ACCESS WITH DISPROT REST API

DisProt can be accessed programmatically via REST API to retrieve a single entry (or region) and to perform large-scale database searches. All API endpoints are available from URL: disprot.org/api/{endpoint_name} . In this support protocol we introduce three different endpoints, the first one can be used to retrieve a single entry, the other two to search entries in the database. Please refer to disprot.org/help#api for all the tables describing identifiers, query parameters and input/output formats mentioned in this Support Protocol.

Necessary resources

Hardware

Laptop or desktop computer. An active and stable internet connection is required.

Software

Python 3.5+, Requests Python library 2.23+.

Input data

No input data are required.

Get a single entity

Users can retrieve a single entity, i.e. a protein entry or one of its manually curated regions, by using its corresponding identifier. The following syntax must be used to retrieve a single entity from DisProt disprot.org/api/{identifier} where the "identifier" must be a valid DisProt ID, DisProt region ID, or a UniProt accession. The query is customizable with various parameters, e.g. file format and release. Here we provide two pieces of code to retrieve a single entry in JSON format (Sample code 1) and in FASTA format (Sample code 2). In Sample code 2 the API version of DisProt is also specified.

Requirements

Python 3.5+, Requests Python library 2.23+.

Sample code 1

```
#!/usr/bin/env python3
import requests

disprot_id = "DP00086"
url = "disprot.org/api/" + disprot_id
resp_json = requests.get(url).json()
print(resp_json)
```

Sample code 2

```
#!/usr/bin/env python3
import requests

disprot_id = "DP00086"
header = {
       'accept-version': '8.0',
       'format':'fasta',
}
url = "disprot.org/api/" + disprot_id
resp_fasta = requests.get(url, headers=header).text
print(resp_fasta)
```

Results

DisProt currently provides three output formats: JSON (default), FASTA and TSV. Due to the inherent limitations of the FASTA and TSV file formats, the JSON format renders the most comprehensive description of intrinsic disorder. The TSV and FASTA files provide details about regions or different types of consensus.

Performing a text search

DisProt provides an extensively customizable search engine. It is possible to perform a free text search or formulate complex queries against combined fields, e.g. organism and UniRef50. The search query is sent to disprot.org/api/search with URL parameters. Note that whitespace and other special characters must be converted into a valid ASCII format; the space is usually replaced with "%20". Multiple search fields can be combined in the same query, by joining them with an AND operator ("&" symbol), e.g. "disprot.org/api/search?organism=homo%20sapiens&name=kinase" returns all the human proteins with "kinase" in the protein name. Given that some fields are interpreted as regular expressions, it is also possible to use the OR operator ("|" symbol). This is the case of the following query, e.g. "disprot.org/api/search?organism=homo%20sapiens|mus %20musculus" which returns both human and mouse entries. The user can choose to customize the output format. Currently available output formats are JSON, FASTA and TSV. By default the endpoint returns the results in JSON, however users can select another format using the "format" field in the parameters or headers. It is possible to use an older version of the API for legacy reasons by specifying the "accept-version" in the URL header of a request. By default the server responds with the latest version of the API.

Requirements

Python 3.5+, Requests Python library 2.23+.

Sample code

```
#!/usr/bin/env python3
import requests

parameters= {
       'ncbi_taxon_id':'9606',
       'format':'tsv'
```

```
}
url = "disprot.org/api/search"
resp_tsv = requests.get(url, params=parameters).text
print(resp_tsv)
```

Results

DisProt returns an object with "data" and "size" fields. "Data" contains a list of entries, and these entry objects are the same described in the previous section. "Size" corresponds to the number of matched entries. Note that when the pagination parameters are provided only the data field is affected whereas the size field always refers to the full query result.

Performing a sequence similarity search

The users can also perform a BLAST sequence similarity search against the database with a POST request to disprot.org/api/blast.

Requirements

Python 3.5+, Requests Python library 2.23+.

Sample code

```
#!/usr/bin/env python3
import requests

data = {
'seq':'KKPLDGEYFTLQIRGRERFEMFRELNEALELKDAQAGKEPGGSRAHSSHLKSKKGQSTSR',
}
header = {
       'accept-version': '8.0'
}
url = "disprot.org/api/blast"
resp_json = requests.post(url, data=data, headers=header).json()
print(resp_json)
```

Results

The output provided is the same available for the text search described above, i.e. JSON (by default), TSV or FASTA. In addition DisProt returns the corresponding "Bit-score", "E-value", "Identity" and "Coverage" as provided by BLAST.

VISUALIZING AND INTERPRETING DISPROT ENTRIES - THE P53 USE CASE

Here we present a use case, human p53 (DisProt entry: DP00086), to explain how to visualize and interpret a DisProt entry page and its annotations. The human p53 entry, also shown in the home page examples, has been recently updated (DisProt release 2020_06) with more than 40 new annotations coming from 15 scientific articles. p53, one of the most well-characterized IDPs, is a tumor suppressor playing a crucial role in several cell functions, such as apoptosis and regulation of DNA repair (Tompa, 2009). p53 is a hub protein, involved in protein-protein interactions with a large number of partners (Uversky & Dunker, 2010). p53 is characterized by the presence of four domains; two of them, the N-terminal transactivation domain (TAD) and the C-terminal tetramerization and regulatory domain, are unstructured as determined by various methods such as NMR, circular dichroism and SAXS (Tompa, 2009; Uversky & Dunker, 2010). Several experimental studies have been carried out in the last two decades that shed light on protein complexes involving p53. Specific short segments of the TAD and the C-terminus domains of p53 with their partners are associated with folding-upon-binding events and are sufficient for interaction-mediated functions (Mészáros et al., 2007).

DisProt entries are annotated by a team of expert curators that aim at collecting all experimental evidence related to disorder available from a publication. In DisProt an entry corresponds to a protein isoform and each IDR annotation is an evidence about its flexible nature or function. The minimal information required to annotate a region in DisProt include reference to the publication (PMID or a DOI), the boundaries of the region (start and end position on the amino acid sequence), the experimental method and the type of information, i.e. an IDP ontology term (*structural state, structural transition, interaction*

partner or *disorder function)*. In order to support annotations, when possible, curators report authors statements as snippets of text from the corresponding publication. Finally, a selected team of reviewers carefully check all annotations, to ensure a high quality standard. Each entry page consists of two main sections. The first provides information about the protein and it includes a feature viewer to visualize DisProt region annotations on the sequence. The second section lists all annotations in a tabular format.

Necessary resources

Hardware

While DisProt works best on laptop or desktop computers, it is also easily accessible from smartphones and tablets. An active and stable internet connection is required.

Software

Internet browser, e.g. Firefox (URL: www.mozilla.org/firefox/), Google Chrome (URL: www.google.com/chrome/), Safari (URL: www.apple.com/safari/).

Input data

Input data are not required.

1. Select the "p53" example, DP00086, from the DisProt home page (disprot.org/).
2. On the top of each entry page the following details are available (Fig. 6): the DisProt identifier (*DP00086*) and protein name (*Cellular tumor antigen p53*), organism (*Homo sapiens*), sequence length (*393*), disorder content (*48.1%*) and cross references with other databases, MobiDB (Piovesan et al., 2018) and UniProt (UniProt Consortium, 2019) (UniProt accession code: P04637).
3. Users can select the release they want to visualize from the "Release" drop-down menu on the top-right of the entry page. All the annotations described in this example correspond to the 2020_06 (latest) release of DisProt.
4. To show/hide ambiguous and/or obsolete regions of an entry, please check/uncheck their corresponding boxes on the top-right of the entry page.
5. The feature-viewer, which can be expanded and collapsed, allows users to visualize regions annotations on the sequence. By default, two tracks are shown, the first showing DisProt annotations and the other including domain data as defined by Pfam (El-Gebali et al., 2019) which provides conserved domain families, and Gene3D (Lewis et al., 2018) which provides globular domains. It is possible to expand the feature viewer to visualize the sub tracks and each disorder evidence annotated for a specific functional or structural aspect. By hovering each region on the sequence viewer, a tooltip provides additional information such as annotated terms, identifiers, cross-references, the name of the curator who annotated the region, the experimental method and the reference supporting that annotation.
6. Users can open ("toggle") the sequence viewer which dynamically highlights amino acids of the selected IDR directly on the protein sequence.
7. It is also possible to select a subset of annotations using the "Filter" box under the sequence viewer.

The bottom section of the entry page lists all DisProt annotations. The N-terminal tail of p53 consists of a transactivation domain (TAD), described in the annotation DP00086r024 (Fig 7.a), spanning from residues 1 to 93 of the protein sequence. The transactivation domain is composed of two subdomains, TAD I and TAD II, and was determined to be unstructured by Fersht et al. (Wells et al., 2008).

The TAD II subdomain is involved in the interaction with the pleckstrin homology (PH) domain from human TFIIH basal transcription factor complex p62 (Okuda & Nishimura, 2014). Binding of the TFIIH PH domain induces a disorder-to-order transition in p53. This interaction plays a crucial role in increasing the affinity of p53 for the transcriptional machinery and might regulate its selectivity for the expression of various genes (Okuda & Nishimura, 2014), therefore supporting the function of p53 as a *molecular recognition effector*. These examples of a p53 interaction (Fig 8), its transition (Fig 9) and the function associated with this binding (Fig 10) are shown in the region annotations available from the p53 entry in DisProt.

PROVIDING FEEDBACK AND SUBMITTING NEW INTRINSIC DISORDER-RELATED DATA

Feedback on site experience and on technical and/or data issues can be submitted using the DisProt feedback form (disprot.org/feedback). In the Feedback page, two tabs are available, "Leave a comment" (Fig. 11) and "Submit a new annotation" (Fig. 12). The first tab allows users to submit a generic feedback, the second to submit a new annotation. DisProt entries are annotated by a team of expert curators, and carefully reviewed by a small team of reviewers. However, the submission of new literature annotations by knowledgeable users is encouraged. Each submitted evidence is reviewed by a team of reviewers and made available in the next release of the database.

Necessary resources

Hardware

Laptop or desktop computer, tablet or smartphone. An active and stable internet connection is required.

Software

Internet browser, e.g. Firefox (URL: www.mozilla.org/en-US/exp/firefox/new/), Google Chrome (URL: www.google.com/chrome/), Safari (URL: www.apple.com/safari/).

Input data

No input data are required.

1. Open a web browser and connect to DisProt: disprot.org/.
2. Click on the "Feedback" button on the top-right of the DisProt bar.
3. Provide your contact information, name and email address, in the corresponding boxes.

Submitting a feedback

4. Select the "Leave a comment" tab.
5. Add a subject of your message in the dedicated field, e.g. "technical issue".
6. Use the "Message" box to add a detailed comment or feedback. The minimum length of the message should be 15 characters.
7. Click on the green "Send" button to send your feedback to the DisProt team.

Submitting a new evidence of intrinsic disorder from the literature

4. Select the "Submit a new annotation" tab.
5. Provide an identifier of the protein you want to annotate, using the "DisProt ID" or the "UniProtKB ACC" boxes. *If the protein is already in DisProt, please provide the DisProt ID, e.g. DP00003. If the protein is not yet annotated in DisProt, please provide its corresponding UniProt identifier, e.g. P03265.*
6. Provide a reference of the publication describing a new evidence of intrinsic disorder for the protein of interest using the "Reference" box. The provided reference must be a valid PubMed ID or DOI of the publication.
7. In the "Experimental method" box add the method used to assess the presence of intrinsic disorder in the publication, e.g. NMR, circular dichroism, small angle X-ray scattering.
8. Add details about the intrinsically disordered region described in the publication. Users can add more than one intrinsically disordered region described in the publication by clicking on the "Add new region" button. To remove a region please click on the "Remove this region".
    a. Provide the boundaries of the intrinsically disordered region: add the start position in the "Start" box and the end position in the "End" box. Region boundaries must correspond to those specified in the publication.
    b. In the "Statement" box please add a sentence from the publication that describes the intrinsically disordered region.
9. Click on the green "Send" button to submit your annotation to the DisProt team.

GUIDELINES FOR UNDERSTANDING RESULTS

In DisProt, a team of expert biocurators manually curated experimental intrinsic disorder data from peer reviewed publications. Each DisProt *entry* corresponds to a UniProt entry, i.e. the canonical sequence or one of its isoforms. An entry consists of a set of manually curated intrinsically disordered regions, each one of them is an *evidence*, together with all the information about its flexible nature. The minimal information included in an evidence is the reference (PMID or DOI) to a scientific publication, an experimental method associated to the IDR, the start and end positions of the region and a disorder aspect associated to the IDR. Four possible *disorder aspects* can be annotated in DisProt, covering the main features of an IDR: the *structural state*, the *structural transition*, its *interaction partner* and the *disorder function*. Each of the aforementioned branches consists of a parent term and its children, e.g. in the "disorder function" branch the parent term *entropic chain* incorporates six child terms, *flexible linker/spacer*, *entropic bristle*, *entropic spring*, *entropic clock*, *entropic spring*, *structural mortar* and *self-transport through channel*. Curators also add *statements*, i.e. sentences from the publication that support the disordered nature of the region or one of its aspects, to provide the users with an exhaustive description of each protein region. A standardized curation effort is one of the main goals of DisProt: in line with this, DisProt curators benefit from a detailed curation manual describing all the

rules to annotate in DisProt and every aspect related to the curation process and of a dedicated ontology of intrinsic disorder-related terms.

ACKNOWLEDGEMENTS:

DisProt is a service of the Italian node of ELIXIR. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie (Grant agreement No. 778247) and by the Italian Ministry of University and Research (MIUR), PRIN 2017 (Grant agreement No. 2017483NH8).

LITERATURE CITED:

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. Journal of Molecular Biology, 215(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Cumberworth, A., Lamour, G., Babu, M. M., & Gsponer, J. (2013). Promiscuity as a functional trait: intrinsically disordered regions as central players of interactomes. Biochemical Journal, 454(3), 361–369. https://doi.org/10.1042/BJ20130545

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., Gibson, T. J., Gutmanas, A., Hancock, J. M., Harrow, J., Higgins, D., Jeffries, C. M., Le Mercier, P., Mészáros, B., Necci, M., … Tosatto, S. C. E. (2019). An intrinsically disordered proteins community for ELIXIR. F1000Research, 8. https://doi.org/10.12688/f1000research.20136.1

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., Sonnhammer, E. L. L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S. C. E., & Finn, R. D. (2019). The Pfam protein families database in 2019. Nucleic Acids Research, 47(D1), D427–D432. https://doi.org/10.1093/nar/gky995

Hatos, A., Hajdu-Soltész, B., Monzon, A. M., Palopoli, N., Álvarez, L., Aykac-Fas, B., Bassot, C., Benítez, G. I., Bevilacqua, M., Chasapi, A., Chemes, L., Davey, N. E., Davidović, R., Dunker, A. K., Elofsson, A., Gobeill, J., Foutel, N. S. G., Sudha, G., Guharoy, M., … Piovesan, D. (2020). DisProt: intrinsic protein disorder annotation in 2020. Nucleic Acids Research, 48(D1), D269–D276. https://doi.org/10.1093/nar/gkz975

Jones, S., & Thornton, J. M. (1996). Principles of protein-protein interactions. Proceedings of the National Academy of Sciences of the United States of America, 93(1), 13–20. https://doi.org/10.1073/pnas.93.1.13

Kobe, B., Guncar, G., Buchholz, R., Huber, T., Maco, B., Cowieson, N., Martin, J. L., Marfori, M., & Forwood, J. K. (2008). Crystallography and protein-protein interactions: biological interfaces and crystal contacts. Biochemical Society Transactions, 36(Pt 6), 1438–1441. https://doi.org/10.1042/BST0361438

Lewis, T. E., Sillitoe, I., Dawson, N., Lam, S. D., Clarke, T., Lee, D., Orengo, C., & Lees, J. (2018). Gene3D: Extensive prediction of globular domains in proteins. Nucleic Acids Research, 46(D1), D435–D439. https://doi.org/10.1093/nar/gkx1069

Mészáros, B., Simon, I., & Dosztányi, Z. (2009). Prediction of protein binding regions in disordered proteins. PLoS Computational Biology, 5(5), e1000376. https://doi.org/10.1371/journal.pcbi.1000376

Mészáros, B., Simon, I., & Dosztányi, Z. (2011). The expanding view of protein-protein interactions: complexes involving intrinsically disordered proteins. Physical Biology, 8(3), 035003. https://doi.org/10.1088/1478-3975/8/3/035003

Mészáros, B., Tompa, P., Simon, I., & Dosztányi, Z. (2007). Molecular principles of the interactions of disordered proteins. Journal of Molecular Biology, 372(2), 549–561. https://doi.org/10.1016/j.jmb.2007.07.004

Mizutani, H., Saraboji, K., Malathy Sony, S. M., Ponnuswamy, M. N., Kumarevel, T., Krishna Swamy, B. S., Simanshu, D. K., Murthy, M. R. N., & Kunishima, N. (2008). Systematic study on crystal-contact engineering of diphthine synthase: influence of mutations at crystal-packing regions on X-ray diffraction quality. Acta Crystallographica. Section D, Biological Crystallography, 64(Pt 10), 1020–1033. https://doi.org/10.1107/S0907444908023019

Okuda, M., & Nishimura, Y. (2014). Extended string binding mode of the phosphorylated transactivation domain of tumor suppressor p53. Journal of the American Chemical Society, 136(40), 14143–14152. https://doi.org/10.1021/ja506351f

Piovesan, D., Tabaro, F., Mičetić, I., Necci, M., Quaglia, F., Oldfield, C. J., Aspromonte, M. C., Davey, N. E., Davidović, R., Dosztányi, Z., Elofsson, A., Gasparini, A., Hatos, A., Kajava, A. V., Kalmar, L., Leonardi, E., Lazar, T., Macedo-Ribeiro, S., Macossay-Castillo, M., … Tosatto, S. C. E. (2017). DisProt 7.0: a major update of the database of disordered proteins. Nucleic Acids Research, 45(D1), D219–D227. https://doi.org/10.1093/nar/gkw1056

Piovesan, D., Tabaro, F., Paladin, L., Necci, M., Micetic, I., Camilloni, C., Davey, N., Dosztányi, Z., Mészáros, B., Monzon, A. M., Parisi, G., Schad, E., Sormanni, P., Tompa, P., Vendruscolo, M., Vranken, W. F., & Tosatto, S. C. E. (2018). MobiDB 3.0: more annotations for intrinsic disorder, conformational diversity and interactions in proteins. Nucleic Acids Research, 46(D1), D471–D476. https://doi.org/10.1093/nar/gkx1071

Ptitsyn, O. B. (1995). Molten globule and protein folding. Advances in Protein Chemistry, 47, 83–229. https://doi.org/10.1016/s0065-3233(08)60546-x

Ptitsyn, O. B., Bychkova, V. E., & Uversky, V. N. (1995). Kinetic and equilibrium folding intermediates. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 348(1323), 35–41. https://doi.org/10.1098/rstb.1995.0043

Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., Wu, C. H., & UniProt Consortium. (2015). UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics , 31(6), 926–932. https://doi.org/10.1093/bioinformatics/btu739
Tompa, P. (2002). Intrinsically unstructured proteins. Trends in Biochemical Sciences, 27(10), 527–533. https://doi.org/10.1016/s0968-0004(02)02169-2
Tompa, P. (2005). The interplay between structure and function in intrinsically unstructured proteins. FEBS Letters, 579(15), 3346–3354. https://doi.org/10.1016/j.febslet.2005.03.072
Tompa, P. (2009). Structure and Function of Intrinsically Disordered Proteins. Taylor & Francis.
UniProt Consortium. (2019). UniProt: a worldwide hub of protein knowledge. Nucleic Acids Research, 47(D1), D506–D515. https://doi.org/10.1093/nar/gky1049
Uversky, V. N., & Dunker, A. K. (2010). Understanding protein non-folding. Biochimica et Biophysica Acta, 1804(6), 1231–1264. https://doi.org/10.1016/j.bbapap.2010.01.017
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., Kim, P. M., Kriwacki, R. W., Oldfield, C. J., Pappu, R. V., Tompa, P., Uversky, V. N., Wright, P. E., & Babu, M. M. (2014). Classification of intrinsically disordered regions and proteins. Chemical Reviews, 114(13), 6589–6631. https://doi.org/10.1021/cr400525m
Wells, M., Tidow, H., Rutherford, T. J., Markwick, P., Jensen, M. R., Mylonas, E., Svergun, D. I., Blackledge, M., & Fersht, A. R. (2008). Structure of tumor suppressor p53 and its intrinsically disordered N-terminal transactivation domain. Proceedings of the National Academy of Sciences of the United States of America, 105(15), 5762–5767. https://doi.org/10.1073/pnas.0801353105

FIGURE LEGENDS:

Figure 1 DisProt main page.

Figure 2 Exploring Disprot. Performing a search with the “Search” boxes on the top-right or top-center of the DisProt home page.

Figure 3 Browse page - Text search. Users can perform advanced text searches, look for specific queries and customize the results of their search.

Figure 4 Browse page - BLAST. Users can perform BLAST searches of a specific protein sequence against the entries available in DisProt.

Figure 5 Download page. All DisProt releases and the IDP ontology are available for downloading in different file formats.

Figure 6 p53 entry page in DisProt. Entry information, feature viewer and sequence viewer and all the features described in steps 1-7 are shown.

Figure 7 Structural state evidence, DP00086r024, describing the unstructured TAD domain of p53.

Figure 8 Interaction partner evidence, DP00086r039, describing the binding of p53 TAD II subdomain to the pleckstrin homology (PH) domain from human TFIIH.

Figure 9 Structural transition evidence, DP00086r040, describing the disorder-to-order transition occurring upon binding of p53 TAD II subdomain to the pleckstrin homology (PH) domain from human TFIIH.

Figure 10 Disorder function evidence, DP00086r041, describing the molecular recognition effector function of p53 TAD II subdomain regulated by binding to the pleckstrin homology (PH) domain from human TFIIH.

Figure 11 Feedback page - Leave a comment. Users can provide feedback on site experience, bugs or issues with data.

Figure 12 Feedback page - Curation mode. Users can submit new pieces of evidence of disorder from literature to the DisProt curators team.