Supplementary material to the Bioinformatics article

"***DIBS: a repository of disordered binding sites mediating interactions with ordered proteins***"

by Eva Schad, Erzsébet Fichó, Rita Pancsa, István Simon, Zsuzsanna Dosztányi and Bálint Mészáros.

The construction of DIBS through the annotation of protein complexes

The basic elements of DIBS are protein complexes, for which constituent chains are all annotated as either ordered or disordered. These structural annotations are derived from various databases (Figure S1), including disorder-specific databases DisProt (Piovesan *et al.*, 2017) and IDEAL (Fukuchi *et al.*, 2012), and the generic structural database PDB (Berman *et al.*, 2000). While the PDB is primarily a database of tertiary/quaternary structure, it can provide evidence of disorder through missing coordinates in X-ray structures or through NMR structures portraying highly variable structural ensembles. These disorder information were used both directly to annotate protein chains shown as 'Proof of disorder: Confirmed', and for homologous proteins as well mirrored in 'Proof of disorder: Inferred from homology' descriptions in DIBS.

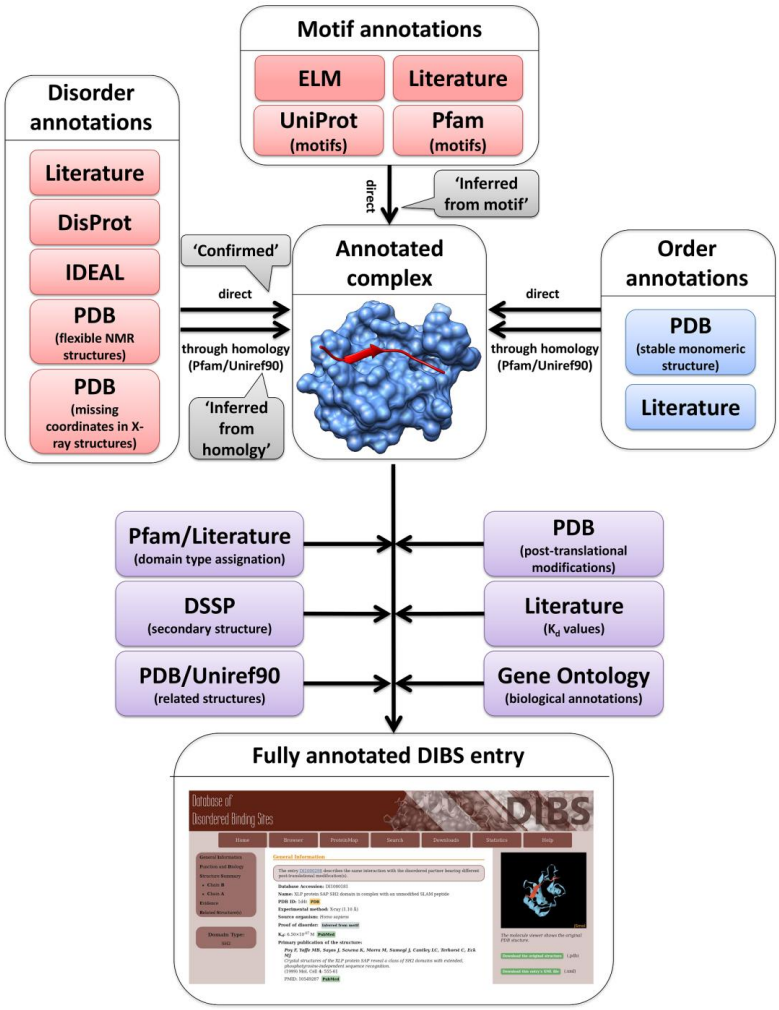

Fig. S1. The sources of annotations in DIBS. Boxes with various colours represent sources of annotations/information. Red boxes mark the source of information about protein disorder, blue boxes represent information about protein order and purple boxes mark all other sources of information not focused on tertiary structure. Black arrows show the direction of the flow of information. Grey caption boxes represent the three types of disorder proof ('Confirmed', 'Inferred from homology' and 'Inferred from motif') connected to various sources of disorder evidence.

The direct evidence of disorder are complemented with indirect proof via the occurrences of short linear motifs (SLiMs). SLiM occurrences were primarily collected from ELM (Dinkel *et al.*, 2016), UniProt (The UniProt Consortium, 2017) and Pfam (Punta *et al.*, 2012). In the case of a SLiM only direct matches with the protein in question were accepted and no homology-transfer of annotations was used.

The main source of order annotations was the PDB. Ordered proteins were required to have a stable solved structure without the bound disordered partner. Similarly to direct disorder proofs, these proofs of order were used both directly and through homology as well using UniRef90 sequence clusters and Pfam objects.

All complexes in DIBS have to have valid direct or indirect proof of disorder for exactly one constituent protein chain and valid direct or indirect proofs of order for all other chains. Protein complexes fulfilling these criteria are included in DIBS and are further annotated with other information about their biological roles and sub-cellular localizations, post-translational modifications, K_d of the interaction (if known), domain type for the ordered partner(s), secondary structures, and a list of similar complexes. For a more complete description of included annotations refer to the 'Help' section of the DIBS server (http://dibs.enzim.ttk.mta.hu/help.php).

For disorder, order and other types of annotations, information from various databases were complemented with extensive manual literature searches. Candidates for disordered proteins were taken from various publications listing several such cases including reviews (e.g. (Wright and Dyson, 2015), (Habchi *et al.*, 2014), (Tompa, 2012), (van der Lee *et al.*, 2014)) and database/method development articles (e.g. (Mészáros *et al.*, 2009), (Mészáros *et al.*, 2017), (Filippakopoulos *et al.*, 2012), (Fichó *et al.*, 2017)), complemented by the expertise of the authors. These articles were thoroughly checked by database curators to find novel examples that were absent from DisProt and IDEAL. For each identified disordered protein the PDB was checked for involvement in protein-protein interactions with ordered partners. Furthermore, found domains interacting with disordered proteins were also checked to see if the PDB contains more DIBS-compatible bound structures of the same domain. In these cases the relevant publications describing the interactions were fully read by the curators to check the disordered status of the partner.

Manual literature searches were also used to extend the level of disorder annotations, i.e. find disorder annotations for proteins with only motif annotations and vice-versa: to find possible known motifs in 'Confirmed' or 'Inferred from homology' entries. These searches were concentrated on the primary publications of the included structures given in the PDB records. These papers were automatically downloaded and checked by pattern matching algorithms for the occurrence of certain key phrases ('disorder', 'unstructur' and 'flex' in the case of search for disorder annotations, and 'motif' for searches for motif occurrences). Papers producing hits were fully read by database curators to eliminate false positive hits and were transformed into annotations in DIBS entries.

A similar text-mining approach was employed in the search for K_d values as well. Primary structure publications were automatically scanned for the occurrence of the term 'Kd' and hits were inspected manually to ensure the inclusion of only reliable binding parameters.

Using the DIBS server

The DIBS server at http://dibs.enzim.ttk.mta.hu/ serves as the main platform to access DIBS. The full database can be downloaded at the 'Downloads' section. Furthermore, data can be accessed online in three different ways shown in Figure S2.

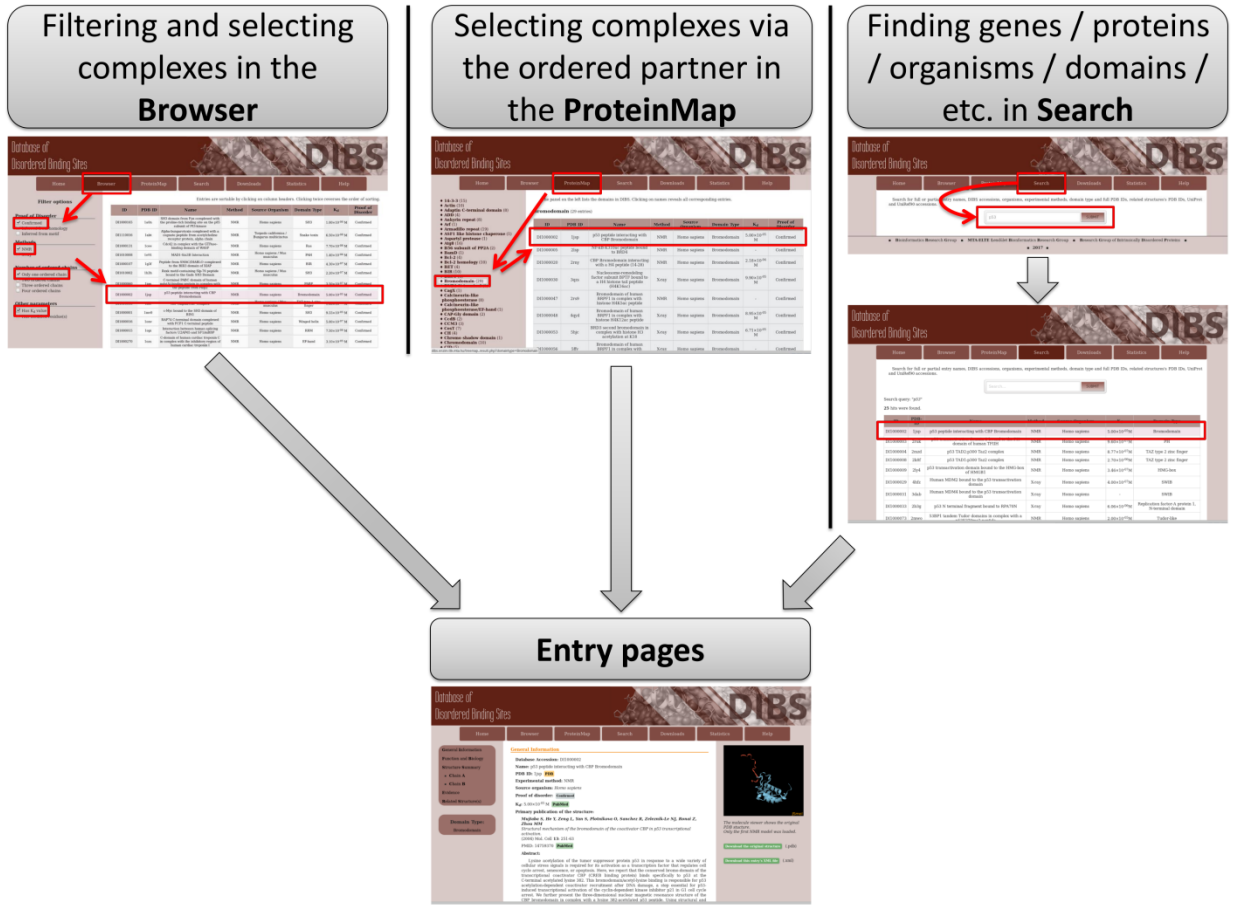

Fig. S2. The three main ways to access data online at the DIBS server. Red boxes mark clickable items while red arrows show the sequentiality of clicks. All three examples using the three different methods lead to the same entry page.

The 'Browser' page lists all entries currently in DIBS. This list can be restricted by applying filters found at the left menu. Complexes can be filtered by proof of disorder, the experimental method used to determine the structure of the complex, the number of ordered partners, or the availability of K_d values.

The 'ProteinMap' offers a way to filter DIBS entries based on the domain type of the ordered partner. Certain ordered domains (such as SH2 or 14-3-3 domains) are known to bind a large number of partner proteins. Selecting specific ordered domain types from the ProteinMap offers a starting point in the inspection of various disordered protein segments binding to a common ordered interactor.

The ‘Search’ field can be used to query DIBS entries matching a given search key. This can include gene/protein names, domain types of the ordered partner, UniProt/UniRef90/PDB and DIBS accessions, organism names, and experimental methods. DIBS also supports a limited way to search by sequence similarity via the Search field. While no input sequence can be submitted to the server, all entries are annotated with UniRef90 cluster names and the search field facilitates the use of these cluster names as search terms. E.g. using the search term ‘UniRef90_Q71DI3’ (the UniRef90 cluster ID for human histone H3.2 and its close homologues) the DIBS server returns complexes including human, murine and drosophila histones as well.

Figure S2 shows the three approaches in action through the example of human p53 bound to the bromodomain of human CBP. Using the Browser, the user can limit the list of presented DIBS entries by selecting only ‘Confirmed’ entries with NMR structures, containing only one ordered protein, and having a K_d value. Then he/she can proceed to select the interaction of interest from this restricted list. The same interaction can be found through the ProteinMap menu, selecting ‘Bromodomain’ as the domain type of interest. As a third option, the user can input ‘p53’ and select the interaction from the results list.

References

Berman,H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Dinkel,H. *et al.* (2016) ELM 2016--data update and new functionality of the eukaryotic linear motif resource. *Nucleic Acids Res.*, **44**, D294–300.

Fichó,E. *et al.* (2017) MFIB: a repository of protein complexes with mutual folding induced by binding. *Bioinformatics*.

Filippakopoulos,P. *et al.* (2012) Histone recognition and large-scale structural analysis of the human bromodomain family. *Cell*, **149**, 214–231.

Fukuchi,S. *et al.* (2012) IDEAL: Intrinsically Disordered proteins with Extensive Annotations and Literature. *Nucleic Acids Res.*, **40**, D507–11.

Habchi,J. *et al.* (2014) Introducing protein intrinsic disorder. *Chem. Rev.*, **114**, 6561–6588.

van der Lee,R. *et al.* (2014) Classification of intrinsically disordered regions and proteins. *Chem. Rev.*, **114**, 6589–6631.

Mészáros,B. *et al.* (2017) Degrons in cancer. *Sci. Signal.*, **10**.

Mészáros,B. *et al.* (2009) Prediction of protein binding regions in disordered proteins. *PLoS Comput. Biol.*, **5**, e1000376.

Piovesan,D. *et al.* (2017) DisProt 7.0: a major update of the database of disordered proteins. *Nucleic Acids Res.*, **45**, D1123–D1124.

Punta,M. *et al.* (2012) The Pfam protein families database. *Nucleic Acids Res.*, **40**, D290–301.

The UniProt Consortium (2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **45**, D158–D169.

Tompa,P. (2012) Intrinsically disordered proteins: a 10-year recap. *Trends Biochem. Sci.*, **37**, 509–516.

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