Project Acronym:	**IDPfun**
Project Full Title:	**Driving functional characterization of intrinsically disordered proteins**
Grant Agreement:	**778247**
Project Duration:	**66 months (01/03/2018 - 31/08/2023)**

Deliverable D4.2

Integration of ID data generated in WP1 and WP2 into MobiDB

Work Package:	**WP4 – IDP translation & deployment**
Lead Beneficiary:	**UNIPD**
Due Date:	**28 Feb 2023 (M60)**
Submission Date:	**28 Feb 2023 (M60)**
Deliverable Type:	**Demonstrator**
Dissemination Level:	**Public**

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

Table of Content

Executive summary

MobiDB (URL: https://mobidb.org/) aggregates disorder annotations derived from the literature and from experimental evidence along with predictions for all known protein sequences. MobiDB generates new knowledge and captures the functional significance of disordered regions by processing and combining complementary sources of information.
The major advances provided in this deliverable are the integration of AlphaFoldDB predictions the re-implementation of the homology transfer pipeline, which expands manually curated annotations by two orders of magnitude and the refinement of the functional characterization of the binding modes for binding regions inside intrinsically disordered regions.

MobiDB integration

MobiDB content

In order to make its content more accessible to the scientific community, MobiDB adopts the concept of “annotation pyramid”. The height of the pyramid represents the annotation quality while the horizontal axis is the coverage of known proteomes. Also, the MobiDB pyramid is staired to indicate different levels of evidence. In Table 1 are reported the number of entries along with the type and source of information for the four levels of the MobiDB pyramid. For each level and feature, MobiDB reports consensus annotations which are combined according to different sets of rules. The complete description of features, sources and consensus strategies are stored in a controlled vocabulary.
Curated annotations are pulled from the corresponding databases processing the data and checking their consistency. Curated databases make use of the tools developed in **WP1, D1.3**. Curated entries are also used as input to infer homology and project their annotation to the rest of UniProtKB sequences (1). The homology pipeline is derived from the work done in **D1.4**. Derived annotations are extracted from PDB structures **(D1.2)** while sequence-based predictions are calculated with MobiDB-lite (2) **(D1.5 & D2.2)** which encapsulate a number of complementary predictors. The subset of disorder features provided by MobiDB-lite are the same provided by InterProScan (3) which propagates its predictions onto several other EBI resources like UniProtKB, InterPro (4) and PDBe-KB (5).

Evidence (size)	Feature	Source
Curated (4,600)	Disorder, LIPs	DisProt
	Disorder, LIPs	IDEAL
	Disorder	Swiss-Prot / UniProtKB
	LIPs	MFIB
	LIPs	DIBS
	LIPs	ELM
	Binding modes	FuzDB
	Conformational diversity	CoDNaS
	LLPS	PhaSePro
	LLPS	PhaSepDB
Derived (59,076)	Disorder, LIPs, Binding modes	* FLIPPER (PDB structures)
	Inter chain interactions	RING (PDB structures)
Homology (458,167)	All curated features	In-house pipeline (curated and UniProtKB sequences)
Predicted (>200 M)	Disorder, LIPs, low complexity, compositional bias, secondary structure, structural rigidity	^ MobiDB-Lite (UniProtKB sequences)
	Disorder, LIPs	AlphaFold-disorder (AlphaFoldDB)

Table 1. Number of entries and annotation source for the four levels of the MobiDB data pyramid (MobiDB release 2022_07). For software sources, the input is indicated in parenthesis. (*) The FLIPPER repository includes MOBI (6) and additional in-house processing scripts to calculate disorder and binding modes features, respectively. (^) The full list of available software integrated into MobiDB-lite is provided in (7).

Homology transfer

In the current version of MobiDB, manually curated annotations are transferred to other proteins based on homology inference. The search for homologous regions is performed starting from a full sequence BLAST alignment against the entire UniProtKB database and applying a filtering procedure in order to minimize the number of false positive instances.
Alignments are performed starting from the full sequences in order to discard non significant matches. The pipeline starts from full sequence alignments but focuses only on alignment fragments corresponding to manually curated regions in MobiDB. The annotation is transferred when very stringent sequence similarity constraints are fulfilled. Specifically, the alignment fragment must cover the 90% of the query sequence (annotated region), gaps must not exceed the 20% of the length of the alignment and the subject (homologous region) must be 80% identical to the query region. In the case of multiple regions being identified on the same target protein, in order to remove overlaps, a greedy algorithm which prioritizes longer regions, is applied. Despite an expansion of two orders of magnitude, the homology transfer for low complexity regions is limited as they are masked by BLAST by default.

AlphaFold and conditional disorder

AlphaFold-2 is the most accurate predictor of protein structures that has been proven to be also effective in identifying intrinsically disordered proteins (8). In MobiDB, AlphaFold predictions are processed in order to extract two alternative definitions of disorder and one definition of linear interacting peptides **(D1.2)**. The first disorder definition is based on the pLDDT score which is a per-residue estimate of the prediction accuracy. In MobiDB residues with a pLDDT lower than 70% are considered disordered. The second definition of disorder is provided by the per-residue relative solvent accessibility (RSA) of the predicted structure, as provided by the DSSP software. The RSA is averaged on a sliding window of 25 residues and positions with an average RSA over 0.58 are considered disordered. The two definitions provide similar results but are complementary at the same time. For example, there are well folded secondary structure elements, e.g. alpha helices, which can be predicted with high confidence (high pLDDT) and at the same time be found inside an extended loopy region disconnected from the rest of the structure and therefore with a high RSA. These regions are likely to undergo conditional (un)folding and can be probably associated with binding events. pLDDT and RSA are therefore combined to also infer LIPs. At the time of writing, all Swis-Prot and model organism proteins are processed and stored in the database, for a total of 1,121,068 entries.

Binding modes

Molecular interactions have a particular significance for IDPs. The structural properties and the amino acid composition of disordered interacting interfaces provides a set of binding modes which are completely different from the canonical lock-and-key mechanism of well structured partners. IDP interactions are mainly provided by electrostatic forces and are entropy-driven resulting in the formation of fuzzy complexes (9). Intrinsically disordered regions (IDRs) can undergo disorder-to-order transitions and fold upon binding, or remain disordered in a partner-bound form. The folding energy of the interaction is compensated by an increase of structural heterogeneity of the rest of the protein. Binding modes of disordered regions refer to the conformational transitions of IDRs upon interacting with specific partners. Some IDRs exhibit context-dependent binding with different partners or cellular conditions. MobiDB aims at collecting as much evidence as possible about the location of binding IDRs in the sequence and about their binding modes.

Binding IDRs in MobiDB are called linear interacting peptides (LIPs), referring to their extended conformation. Similarly to disorder evidence, in MobiDB there are different levels of evidence (annotation confidence) and different features, the binding modes, that can be associated with a LIP. In Table 2 are shown the types of annotations currently in MobiDB. Whenever available the binding mode is provided, otherwise the region is annotated simply as a LIP. Despite curated databases capturing different binding specificity or subclass, e.g. the Eukaryotic Linear Motif (ELM) database annotates short linear motifs (SLiMs) (10), only FuzDB annotations are associated with a binding mode. FuzDB describes “fuzzy” complexes that remain disordered in the bound state, “disorder-to-disorder” transitions. Other annotations relative to binding modes are provided by an internal pipeline that derives this information from PDB structures by comparing disordered residues in free and bound form,

as described in (11), and using the RING software to detect intermolecular interactions (12). PDB complexes are also processed by the FLIPPER (13) classifier that extracts generic LIP annotations looking at the geometrical and physicochemical properties of the structure **(D1.5)**. Large scale LIPs predictions from sequence are provided by ANCHOR (14), while AlphaFold structures are used to derive LIPs likely to be associated with conditional folding and binding modes (see AlphaFold and conditional disorder paper (8)).

Evidence	Feature	Source	Proteins	Description
Curated	LIP	DisProt, IDEAL	970	All LIP types
		ELM	75	Short Linear Motifs (SLiMs)
		DIBS	498	LIP interacting with structure
		MFIB	246	LIP interacting with LIP
	Binding mode	FuzDB	328	Fuzzy complexes
Derived	LIP	FLIPPER	10,728	All LIP types
	Binding mode	RING	16,606	Structural transition
Predicted	LIP	AlphaFold	991,606	Structural transition
		ANCHOR	>130 M	All LIP types

Table 2. Binding knowledge provided by MobiDB.

References

1. The UniProt Consortium (2021) UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.*, **49**, D480–D489.
2. Necci,M., Piovesan,D., Clementel,D., Dosztányi,Z. and Tosatto,S.C.E. (2020) MobiDB-lite 3.0: fast consensus annotation of intrinsic disorder flavors in proteins. *Bioinformatics*, 10.1093/bioinformatics/btaa1045.
3. Jones,P., Binns,D., Chang,H.-Y., Fraser,M., Li,W., McAnulla,C., McWilliam,H., Maslen,J., Mitchell,A., Nuka,G., *et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinforma. Oxf. Engl.*, **30**, 1236–1240.
4. Blum,M., Chang,H.-Y., Chuguransky,S., Grego,T., Kandasaamy,S., Mitchell,A., Nuka,G., Paysan-Lafosse,T., Qureshi,M., Raj,S., *et al.* (2021) The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.*, **49**, D344–D354.
5. Consortium,Pdb.-K., Varadi,M., Anyango,S., Armstrong,D., Berrisford,J., Choudhary,P., Deshpande,M., Nadzirin,N., Nair,S.S., Pravda,L., *et al.* PDBe-KB: collaboratively defining the biological context of structural data. *Nucleic Acids Res.*, 10.1093/nar/gkab988.
6. Martin,A.J.M., Walsh,I. and Tosatto,S.C.E. (2010) MOBI: a web server to define and

visualize structural mobility in NMR protein ensembles. *Bioinformatics*, **26**, 2916–2917.
7. 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.
8. Piovesan,D., Monzon,A.M. and Tosatto,S.C.E. Intrinsic Protein Disorder and Conditional Folding in AlphaFoldDB. *Protein Sci.*, **n/a**, e4466.
9. Piovesan,D., Arbesú,M., Fuxreiter,M. and Pons,M. (2022) Editorial: Fuzzy Interactions: Many Facets of Protein Binding. *Front. Mol. Biosci.*, **9**.
10. Kumar,M., Michael,S., Alvarado-Valverde,J., Mészáros,B., Sámano-Sánchez,H., Zeke,A., Dobson,L., Lazar,T., Örd,M., Nagpal,A., *et al.* (2022) The Eukaryotic Linear Motif resource: 2022 release. *Nucleic Acids Res.*, **50**, D497–D508.
11. Miskei,M., Horvath,A., Vendruscolo,M. and Fuxreiter,M. (2020) Sequence-Based Prediction of Fuzzy Protein Interactions. *J. Mol. Biol.*, **432**, 2289–2303.
12. Clementel,D., Del Conte,A., Monzon,A.M., Camagni,G.F., Minervini,G., Piovesan,D. and Tosatto,S.C.E. (2022) RING 3.0: fast generation of probabilistic residue interaction networks from structural ensembles. *Nucleic Acids Res.*, 10.1093/nar/gkac365.
13. Monzon,A.M., Bonato,P., Necci,M., Tosatto,S.C.E. and Piovesan,D. (2021) FLIPPER: Predicting and Characterizing Linear Interacting Peptides in the Protein Data Bank. *J. Mol. Biol.*, **433**, 166900.
14. Dosztányi,Z., Mészáros,B. and Simon,I. (2009) ANCHOR: web server for predicting protein binding regions in disordered proteins. *Bioinformatics*, **25**, 2745–2746.

Availability

The MobiDB website is available at URL: https://mobidb.org

The AlphaFold-disorder script is available for download at URL: https://github.com/BioComputingUP/AlphaFold-disorder

The FLIPPER repository which includes MOBI and additional in-house processing scripts to calculate disorder and binding modes features, is available at URL: https://github.com/BioComputingUP/FLIPPER