BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 15 2012, pages 2080–2081
doi:10.1093/bioinformatics/bts327

*Structural bioinformatics*

Advance Access publication June 1, 2012

# MobiDB: a comprehensive database of intrinsic protein disorder annotations

Tomás Di Domenico, Ian Walsh, Alberto J.M. Martin and Silvio C.E. Tosatto*

Department of Biology, University of Padova, Viale G. Colombo 3, 35131 Padova, Italy

Associate Editor: Anna Tramontano

**ABSTRACT**

**Motivation:** Disordered protein regions are key to the function of numerous processes within an organism and to the determination of a protein's biological role. The most common source for protein disorder annotations, DisProt, covers only a fraction of the available sequences. Alternatively, the Protein Data Bank (PDB) has been mined for missing residues in X-ray crystallographic structures. Herein, we provide a centralized source for data on different flavours of disorder in protein structures, MobiDB, building on and expanding the content provided by already existing sources. In addition to the DisProt and PDB X-ray structures, we have added experimental information from NMR structures and five different flavours of two disorder predictors (ESpritz and IUpred). These are combined into a weighted consensus disorder used to classify disordered regions into flexible and constrained disorder. Users are encouraged to submit manual annotations through a submission form. MobiDB features experimental annotations for 17 285 proteins, covering the entire PDB and predictions for the SwissProt database, with 565 200 annotated sequences. Depending on the disorder flavour, 6–20% of the residues are predicted as disordered.

**Availability:** The database is freely available at http://mobidb.bio.unipd.it/.

**Contact:** silvio.tosatto@unipd.it

Received on March 19, 2012; revised on May 23, 2012; accepted on May 30, 2012

## 1 INTRODUCTION

During the last decade, strong evidence has surfaced indicating that many proteins function in a natively unfolded or intrinsically disordered state (Dunker *et al*., 2008; Wright and Dyson, 1999). These regions have been shown to play important roles in various biological processes (Tompa, 2010). The amount of disorder within a proteome seems to correlate with the complexity of the organism, especially in eukaryotes (Ward *et al*., 2004). The existence of different flavours of disorder has been proposed (Vucetic *et al*., 2003), and disordered regions have been categorized according to their function with a suggested coupling between disorder conservation and protein function (Bellay *et al*., 2011; Schlessinger *et al*., 2011).

The main repository for experimentally determined disorder is the DisProt database (Sickmeier *et al*., 2007), containing manually curated information on currently ca. 650 proteins from the literature. Although invaluable as a gold standard, DisProt represents only a fraction of the known protein sequences posing a bottleneck for large-scale analysis of intrinsic protein disorder. Many prediction methods have long resorted to considering the lack of coordinates in X-ray protein structures as a proxy for intrinsic disorder (Walsh *et al*., 2011; Ward *et al*., 2004). This increases the number of available sequences by an order of magnitude for mostly short disordered segments. Recently, our group has also developed a method to define intrinsic disorder by looking at mobile regions in NMR structures (Martin *et al*., 2010). Herein we describe MobiDB, a centralized resource for disorder annotation in protein sequences.

## 2 IMPLEMENTATION

MobiDB is a relational PostgreSQL database consisting of 11 tables. The data are divided into two subsets: MobiDB-xp, containing only proteins with experimental annotation and MobiDB-full, for proteins with predictions. Annotations are extracted from different sources, currently yielding eight-different flavours. The PDB-X-ray data are obtained by considering as disordered residues whose $C_\alpha$ atoms are missing from X-ray crystallographic structures deposited in the PDB (Berman *et al*., 2007). The novel PDB-NMR is generated by processing NMR structures in the PDB with MOBI (Martin *et al*., 2010) and DisProt data (Sickmeier *et al*., 2007) are obtained directly. Predictions are obtained by running ESpritz (Walsh *et al*., 2012) (long, X-ray and NMR) and IUPred (Dosztanyi *et al*., 2005) (short and long) on all SwissProt sequences. Sensitivity and specificity values of each predictor on a common benchmark can be found online. Sequences are linked to UniProt (The UniProt Consortium, 2011) and Pfam (Finn *et al*., 2010) through SIFTS (Velankar *et al*., 2005) for PDB structures and DisProt. A consensus disorder score assigns higher weights to experimental annotations over predictions (see online documentation). Disorder is divided into constrained and flexible based on conservation (Bellay *et al*., 2011). Secondary structure in PDB files is identified with DSSP (Kabsch and Sander, 1983). Manual data curation is also supported and users are encouraged to submit annotations through a feedback submission form.

## 3 USAGE

MobiDB was designed with two main scenarios in mind. First, a user wishes to analyse a particular protein of interest and dynamically access all the available disorder information, with the option to generate (and download) a consensus annotation. Second, the user would like to obtain a dataset of disorder information for a protein ensemble with certain characteristics, downloading it for offline usage and analysis with other tools. MobiDB offers two options

*To whom correspondence should be addressed.

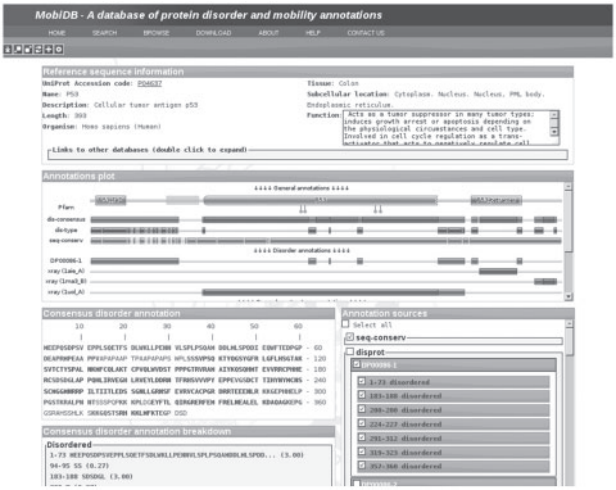


**Fig. 1.** Sample MobiDB output for human p53. The top part contains the UniProt description and database links. The sequence plot (central part) summarizes the disorder information graphically, showing the protein sequence horizontally annotated with Pfam domains and the different disorder flavours. Experimental data are shown in stronger colours (ordered in blue and disordered in red) than predictions. Consensus disorder (blue to red colour gradient) and conservation annotations are also shown. The detailed annotations (bottom part) allow the dynamic selection of annotating sequences and show the relevant sequence stretches

to access the information. The user may either browse the different MobiDB-xp flavours (PDB-X-ray, PDB-NMR or DisProt) or use the search function. The latter offers three options: by identifier, standard and BLAST (Altschul *et al*., 1997). For a full explanation please refer to the online documentation. After selecting a browse option or performing a search, the user will be presented with the results page. In this page, it is possible to either select a single entry and proceed to the protein visualization interface or to generate a dataset containing disorder annotations for all selected proteins. This dataset will consist of two FASTA formatted files for each protein: one containing an alignment of the reference sequence and all the annotating sequences and the second containing annotations associated to these sequences.

The protein visualization interface (Fig. 1) was designed as an annotation sandbox for dynamical protein annotation. The interface is composed of a variety of widgets or boxes that can be dragged, expanded or collapsed, allowing for the optimization of the available workspace. The 'reference sequence information' widget displays data for the chosen reference sequence from UniProt. The 'annotation sources' widget allows selecting or deselecting annotating sequences, and/or their corresponding regions. The 'annotations plot' widget offers a graphical representation of the reference sequence and the chosen annotating sequences, while also displaying Pfam and secondary structure annotations (where available). The 'dynamic annotation' widget displays the colour coded sequence for the reference protein, according to whether a region is annotated as ordered or disordered. A second set of three colours for predicted disorder annotations is provided (in lighter shades). Consensus disorder predictions are provided together with a classification into flexible and constrained regions (Bellay *et al*., 2011).

As an example, we show the annotations for the human p53 tumour suppressor protein in Figure 1. p53 contains structured tetramerization and core domains linked together and flanked by intrinsically disordered regions. The structure of p53 (or lack thereof) has been widely studied, and a comprehensive model has been built (Wells *et al*., 2008). The MobiDB entry for p53 summarizes this situation well (Fig. 1).

MobiDB provides the means to obtain disorder annotations for an extensive set of proteins as a centralized and up-to-date source of information on various available disorder flavours. We are planning on providing manual annotations and integrating more data generated from other predictors to better characterize different disorder flavours and their functional implications.

*Funding*: The University of Padova (CPDA098382, CPDR097328), FIRB Futuro in Ricerca (RBFR08ZSXY) and Cariplo (2017/0724) to S.T.

*Conflict of Interest*: none declared.

## REFERENCES

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Bellay,J. *et al.* (2011) Bringing order to protein disorder through comparative genomics and genetic interactions. *Genome Biol.*, **12**, R14.

Berman,H. *et al.* (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.*, **35**, D301–D303.

Dosztanyi,Z. *et al.* (2005) The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J. Mol. Biol.*, **347**, 827–839.

Dunker,A.K. *et al.* (2008) The unfoldomics decade: an update on intrinsically disordered proteins. *BMC Genom.*, **9** (Suppl 2), S1.

Finn,R.D. *et al.* (2010) The Pfam protein families database. *Nucleic Acids Res.*, **38**, D211–D222.

Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Martin,A.J. *et al.* (2010) MOBI: a web server to define and visualize structural mobility in NMR protein ensembles. *Bioinformatics*, **26**, 2916–2917.

Schlessinger,A. *et al.* (2011) Protein disorder—a breakthrough invention of evolution? *Curr. Opin. Struct. Biol.*, **21**, 412–418.

Sickmeier,M. *et al.* (2007) DisProt: the Database of Disordered Proteins. *Nucleic Acids Res.*, **35**, D786–D793.

The UniProt Consortium (2011) Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res.*, **39**, D214–D219.

Tompa,P. (2010) *Structure and Function of Intrinsically Disordered Proteins*. CRC Press/Taylor and Francis Group, Boca Raton, FL.

Velankar,S. *et al.* (2005) E-MSD: an integrated data resource for bioinformatics. *Nucleic Acids Res.*, **33**, D262–D265.

Vucetic,S. *et al.* (2003) Flavors of protein disorder. *Proteins*, **52**, 573–584.

Walsh,I. *et al.* (2012) ESpritz: accurate and fast prediction of protein disorder. *Bioinformatics*, **28**, 503–509.

Walsh,I. *et al.* (2011) CSpritz: accurate prediction of protein disorder segments with annotation for homology, secondary structure and linear motifs. *Nucleic Acids Res.*, **39**, W190–W196.

Ward,J.J. *et al.* (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.*, **337**, 635–645.

Wells,M. *et al.* (2008) Structure of tumor suppressor p53 and its intrinsically disordered N-terminal transactivation domain. *Proc. Natl. Acad. Sci. U S A*, **105**, 5762–5767.

Wright,P.E. and Dyson,H.J. (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.*, **293**, 321–331.