Published online 31 October 2014 *Nucleic Acids Research, 2015, Vol. 43, Database issue* ***D315–D320***
doi: 10.1093/nar/gku982

MobiDB 2.0: an improved database of intrinsically disordered and mobile proteins

Emilio Potenza, Tomás Di Domenico, Ian Walsh and Silvio C.E. Tosatto*

Department of Biomedical Sciences, University of Padua, 35131 Padova, Italy

Received August 15, 2014; Revised October 2, 2014; Accepted October 03, 2014

ABSTRACT

MobiDB (http://mobidb.bio.unipd.it/) is a database of intrinsically disordered and mobile proteins. Intrinsically disordered regions are key for the function of numerous proteins. Here we provide a new version of MobiDB, a centralized source aimed at providing the most complete picture on different flavors of disorder in protein structures covering all UniProt sequences (currently over 80 million). The database features three levels of annotation: manually curated, indirect and predicted. Manually curated data is extracted from the DisProt database. Indirect data is inferred from PDB structures that are considered an indication of intrinsic disorder. The 10 predictors currently included (three ESpritz flavors, two IUPred flavors, two DisEMBL flavors, GlobPlot, VSL2b and JRONN) enable MobiDB to provide disorder annotations for every protein in absence of more reliable data. The new version also features a consensus annotation and classification for long disordered regions. In order to complement the disorder annotations, MobiDB features additional annotations from external sources. Annotations from the UniProt database include post-translational modifications and linear motifs. Pfam annotations are displayed in graphical form and are link-enabled, allowing the user to visit the corresponding Pfam page for further information. Experimental protein–protein interactions from STRING are also classified for disorder content.

INTRODUCTION

Proteins have been known to exist in an equilibrium between an unfolded and folded state at least since Anfinsen's experiments on denaturation. The existence of an unfolded, or disordered, state has long been considered temporary, due to the protein still having to adopt its final conformation. In this view, mobility of the protein structure was seen as a localized phenomenon, where protein structure determines function and local flexibility is limited to helping the protein achieve its function. This paradigm has been challenged by the collection of hundreds of proteins where function is determined by non-folding regions which play vital biological roles (1,2). Flexible segments lacking a unique native structure, known as intrinsic disordered regions, are widespread in nature, especially in eukaryotic organisms (3,4). The size of disordered regions can be short, long or even encompass entire proteins and their non-enzymatic functions include regulation, protein–DNA/RNA interactions and molecular recognition to name a few, for a recent review see e.g. (5).

One of the first repositories for experimentally determined disorder was the DisProt database (6), containing manually curated information on currently 694 proteins. More recently, the IDEAL database (7) was developed, which annotates 446 proteins with disorder and other interesting properties by scanning the literature. Although DisProt and IDEAL are invaluable as an experimental gold standard, they both represent only a fraction of the sequences in nature, posing a bottleneck for large-scale understanding of the disorder phenomenon. Experimental *in vitro* techniques such as nuclear magnetic resonance (NMR) and x-ray crystallography detect disorder with difficulty in particular for long regions and entire proteins. With currently around 100 000 NMR and x-ray structures, the Protein Data Bank (PDB) (8) nevertheless provides a rich source of indirect experimental disorder. Missing residues in x-ray crystallographic structures in particular have become the *de facto* standard proxy to infer disorder (1,6,9–10). Only more recently have mobile regions in NMR structures started to be used to infer disorder (11), although it is not entirely clear how this relates to either missing x-ray regions or flexible loops. Due to the difficulty in determining disorder experimentally, a plethora of predictors were created over the last 15 years. Many are quite accurate, as shown at the recent Critical Assessment of techniques for protein Structure Prediction (CASP-10) (12) and a large-scale assessment of disorder predictors (10). Biophysical methods (13,14) derive pseudo-energy functions from residue pairings in rigid structures (i.e. non-disorder). Machine learning, especially neural networks, has been widely used to predict protein disorder (3,15–18). Many predictors try to capture quite diverse disorder flavors, e.g. ESpritz (15) can predict mobile

*To whom correspondence should be addressed. Tel: +39 049 8276269; Fax: +39 049 8276260; Email: silvio.tosatto@unipd.it

Table 1. Disorder sources consensus definition matrix

DisProt	PDB	Predictors	Consensus
Disorder	**Disorder**	***Any***	**Disorder**
Disorder	Structure	*Any*	Ambiguous
Disorder	Ambiguous	*Any*	Ambiguous
Structure	Disorder	*Any*	Ambiguous
Structure	Structure	*Any*	Structure
Structure	Ambiguous	*Any*	Ambiguous
Ambiguous	*Any*	*Any*	Ambiguous
None	**Disorder**	***Any***	**Disorder**
None	Structure	*Any*	Structure
None	Ambiguous	*Any*	Ambiguous
None	***None***	**Disorder**	**Disorder (LC)**
None	*None*	Structure	Structure (LC)

Each possible annotation scenario is listed for for the three data sources (DisProt, PDB, predictors) together with its consensus annotation. Ambiguous is used for residues with conflicting annotations warranting further investigation, which may be due to folding upon binding events. LC means low confidence. Combinations yielding structure as consensus are underlined and those for disorder are shown in bold. Sources which are not contributing to the consensus are shown in italics.

NMR regions and DisEMBL (17) loop regions with high B-factor (high flexibility). Predictions can increase the number of annotated sequences to millions but they must be fast to process many gigabytes of data and keep pace with data expansion. Despite earlier interest in proteome-scale disorder predictions (3), DICHOT (19) is probably the first public database to provide predictions for the human proteome (ca. 20 000 proteins). MobiDB (20), initially limited to ca. 450 000 SwissProt sequences, was the first published database to contain a mixture of experimental data and a consensus prediction approach to annotate as many sequences as possible with intrinsic disorder. A similar large-scale database, D2P2 (21), was published somewhat later to provide consensus predictions for ca. 10 million sequences from fully-sequenced genomes. The new version of MobiDB 2.0 improves over its predecessor in terms of coverage and molecular annotations. It is cross-linked from UniProt, covering all of its protein sequences, presently annotating over 80 million sequences from thousands of organisms.

DATABASE DESCRIPTION

Data sources

MobiDB is designed in three layers (in order of quality): manual curation, indirect experimental PDB information and predictions. Its data sources are essentially four: DisProt, PDB-NMR, PDB-xray and predictors. The highest quality data is currently extracted from the DisProt database (6), a central repository manually curated for structure-function annotations associated with protein intrinsic disorder. PDB-NMR disorder, or rather mobility, is generated by processing NMR structures in the PDB with Mobi (11). Deposited files of NMR experiments for protein structure resolution often contain multiple models. By calculating the differences between the positions of each model's residues, the degree in which positions change can be measured, which is interpreted as a measure of how mobile or disordered a protein is. Indirect data is also inferred from missing residues in PDB-xray structures by considering as disordered residues whose Cα atoms are missing from x-ray crystallographic structures deposited in the PDB (8). Furthermore, every sequence in MobiDB is linked to UniProt (22), PDB (8) and Pfam (23) through SIFTS (24). MobiDB also includes secondary structure derived from PDB files using DSSP (25). Pfam annotations are displayed in graphical form and are link-enabled, allowing the user to visit the corresponding Pfam page for further information. Low-complexity regions predicted with SEG (26) and Pfilt (27) are included, as it is thought that low sequence complexity correlates with intrinsic disorder (28,29). Protein–protein interactions are incorporated from STRING (30) by considering only interactions of high accuracy with database or experimental evidence. Functional information from UniProt, e.g. post-translational modifications and binding sites (among others), are also assigned to residues.

Disorder predictors

MobiDB uses three biophysical predictors (IUPred-short (14), IUPred-long (14), Globplot (13)) and seven machine learning predictors (DisEMBL-465 , DisEMBL-HL , Espritz-DisProt (15), Espritz-NMR (15), Espritz-xray (15), JRONN (16) and VSL2b (18)). All predictors are chosen for their speed (<10 s per protein). A consensus prediction is formed by applying a majority vote on the 10 predictors when there is no high quality information from NMR, x-ray or DisProt.

Combining experimental data

The core of MobiDB is shown in the section 'Sequence annotations' where all the data are collected to form a global consensus. The first line of information is dedicated to 'long disorder' consensus and related percentage of residues, as well as the last line is dedicated to 'predictor' consensus as already described. The second line of information 'Disorder Sources' contains the overall representation of disorder that came from the union of DisProt, PDB and predictor consensus. Basically, for each source of information a consensus has been calculated in three possible states: structure, disorder and ambiguous. These are then merged in an overall consensus, using the logic described in Table 1. Simply put, the consensus assigns disorder and structure only when no contradictions are found and ambiguous otherwise.

Your search for "name:"P53" AND organism:"human"" returned "262" results. ×

#	% LD	Entry ID	Length	Protein Name	Organism
1	10.18 %	P04637	393	Cellular tumor antigen p53	Homo sapiens
2	17.73 %	Q13625	1128	Apoptosis-stimulating of p53 protein 2	Homo sapiens
3	15.04 %	Q9UQB8	552	Brain-specific angiogenesis inhibitor 1-associated protein 2	Homo sapiens
4	0.000 %	Q96FX8	193	p53 apoptosis effector related to PMP-22	Homo sapiens
5	43.14 %	Q15648	1581	Mediator of RNA polymerase II transcription subunit 1	Homo sapiens
6	03.44 %	O94776	668	Metastasis-associated protein MTA2	Homo sapiens
7	09.28 %	O15350	636	Tumor protein p73	Homo sapiens
8	46.79 %	Q9NS56	1045	E3 ubiquitin-protein ligase Topors	Homo sapiens
9	62.02 %	Q12888	1972	Tumor suppressor p53-binding protein 1	Homo sapiens
10	29.74 %	Q9BUR4	548	Telomerase Cajal body protein 1	Homo sapiens
11	0.000 %	A1A5B4	782	Anoctamin-9	Homo sapiens
12	0.000 %	Q15051	598	IQ calmodulin-binding motif-containing protein 1	Homo sapiens
13	05.92 %	O15151	490	Protein Mdm4	Homo sapiens
14	0.000 %	Q53FA7	332	Quinone oxidoreductase PIG3	Homo sapiens
15	26.33 %	Q96KQ4	1090	Apoptosis-stimulating of p53 protein 1	Homo sapiens
16	45.08 %	Q9BXH1	193	Bcl-2-binding component 3	Homo sapiens
17	25.00 %	Q9H305	208	Cell death-inducing p53-target protein 1	Homo sapiens
18	0.000 %	O14683	189	Tumor protein p53-inducible protein 11	Homo sapiens
19	0.000 %	Q9NUG6	133	p53 and DNA damage-regulated protein 1	Homo sapiens
20	10.83 %	Q96A56	240	Tumor protein p53-inducible nuclear protein 1	Homo sapiens

First Previous 1 2 3 4 5 6 7 8 9 10 Next Last

Figure 1. Search results page. In this example the keyword 'P53' in organism 'human' is searched and the first 20 results (out of 262) are shown. Long disorder (% LD) coloring is as follows: none (white), low (green), medium (yellow), high (red) and full (black, not shown). Default sorting is by UniProt results, but can be changed by clicking on% LD or length.

Long disorder and classification

Proteins with long disorder regions are more frequent in higher Eukaryotes and known to have specific functions (3,5) as well as being associated with human diseases such as cancer (31). The prediction consensus is also optimized for detection of long disordered regions by optimizing the agreement factor (number of predictors agreeing $\geq$75%) and a regular expression on long regions >20 consecutive amino acids. Optimization is achieved using a grid search and small disordered regions (<10 consecutive residues) are removed. The percentage of disordered residues in long regions is calculated to allow an easier search for interested users. Three classes are defined: high (>30%), medium (15–30%) and low (0–15%) long disorder percentage. Thresholds have been optimized for three uniform sequence subsets over a reduced test set with 10 million proteins.

Implementation

MobiDB was designed with a multi-tier architecture, as previously used in RepeatsDB (32), using separate modules for data management, data processing and presentation functions. To simplify development and maintenance, all tiers handle the common JSON (JavaScript Object Notation) format, thereby eliminating the need for data conversion. The MongoDB database engine is used for data storage and Node.js as middleware between data and pre-

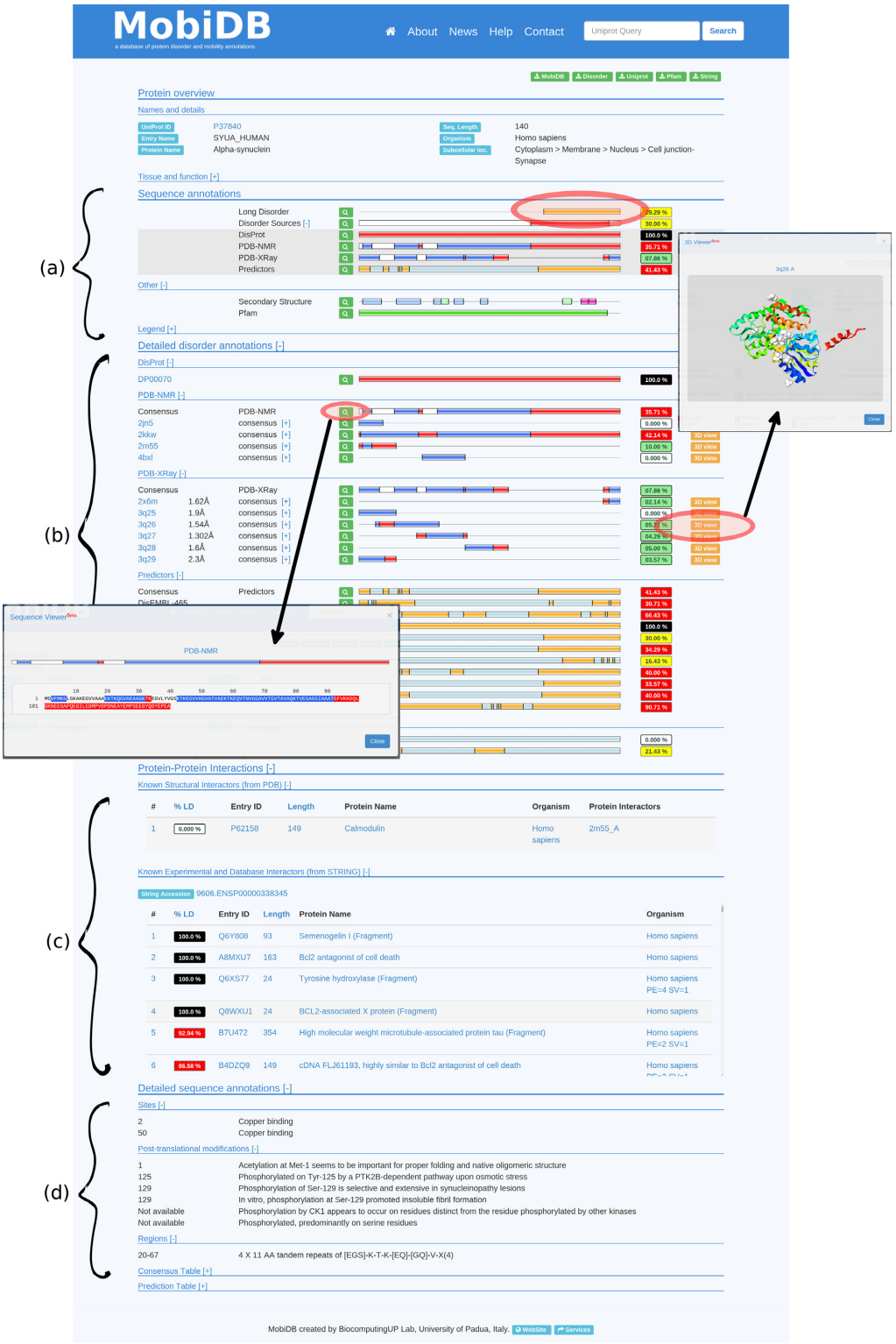

Figure 2. Sequence annotations for alpha-synuclein (UniProt entry: P37840). (**a**) Overview disorder annotations combining DisProt, NMR, x-ray and predictors are shown. The highlighted red circle shows the experimentally determined and predicted long disordered region. Other information includes secondary structure and Pfam domains. Each of these annotations can be downloaded by clicking on the corresponding green button on the top right side of the page. (**b**) Detailed disorder annotation showing experimental (DisProt, NMR and x-ray) and predicted disorder (10 predictors). For each entry, it is possible to view the detailed sequence annotation by clicking on the green magnifying glass icon (see red circle and left inset). Where available, the 3D structure can be visualized to inspect interesting protein regions (see red circle and right inset). The red circle highlights the only known complete structure alpha-synuclein structure (PDB entry 2kkw). (**c**) Known protein–protein interactions deduced from PDB files and STRING are shown in analogy to the search results page, with color-coded long disorder percentage, length, protein name and organism. (**d**) Functional sequence features from UniProt, including binding sites, post-translational modifications and sequence regions.

sentation. The Angular.js framework and Bootstrap library provide the overall look-and-feel. Additional information is added to entries by querying the Uniprot, PDB and Pfam web services. MobiDB offers users both graphical web interface access and exposes its resources through RESTful web services, using the Restify library for Node.js from URL: http://mobidb.bio.unipd.it/. A detailed web service usage guide is available online. MobiDB was designed to be synchronized with UniProt releases with MobiDB updating its own data accordingly, and is already included in UniProt cross-references since the January 2014 release.

USING MobiDB

In the main usage scenario the user is able to analyze a particular protein in terms of its mobility and disorder information either by directly accessing the entry page with an UniProt accession number or by browsing directly from UniProt to our web-site. MobiDB also offers the capability to search the database directly through an advanced query syntax with a complete list of supported query fields for searching specific data (a full explanation can be found in the online documentation). After selecting a query and performing a search, the user will be presented with the results page. Figure 1 shows the results page after searching for 'P53' in organism 'human'. In this page, it is possible to either select a single entry and proceed to the protein visualization interface or sort the results. Sorting for better selection criteria is possible either on protein length or percentage of residues in long disordered regions. In order to understand the disorder phenomenon better three classes of long disorder are defined. Low, medium and high disorder are colored green, yellow and red respectively, with the additional special cases of none (white) and full disorder (black) (see Figure 1). Additional information such as the basic UniProt descriptions and organism are also displayed to aid selection.

The sequence visualization interface is shown in Figure 2 for alpha-synuclein, a protein involved in neurodegenerative disorders which is not yet well understood. The page is composed of a variety of boxes and sections that can be collapsed to optimize usage of the available workspace. Starting from the top right corner (Figure 2a), five download buttons are available for retrieving disordered row data and the other related annotations. In the 'Protein overview' box the user can find a basic description of the sequence, like Uniprot ID, protein name, organisms and so on. The main annotations located inside 'Sequence annotations' (Figure 2a), are displayed as bars by combining the original data sources. By clicking on the green magnifying glass button next to each annotation, it is possible to open a more detailed sequence viewer. The bars titled Disorder Sources, DisProt, PDB-NMR and PDB-xray are defined in the section 'Combining experimental data'. While the prediction bars Predictors and Long Disorder are defined in 'Disorder Predictors' and 'Long Disorder and Classification' sections respectively. Other bars give a more comprehensive picture of the protein, displaying Pfam and secondary structure annotations. More detail is also shown on the visualization page. Figure 2b shows the detailed overview of the raw data, i.e. Disport, PDB-NMR, PDB-xray and Predictors in the section 'Detailed disorder annotations'. Where a PDB is available, the user can visualize the protein structure in 3D, chain by chain or in the entire complex. Scrolling down the page, known interacting proteins from the PDB and STRING are classified by disorder content (see Figure 2c). Last but not least, relevant functional features provided by UniProt, such as post-translational modifications, binding site residues and low complexity regions, can be found at the bottom of the page (see Figure 2d). For a complete summary of MobiDB 2.0 improvements over the previous version see Supplementary Table S1. All the different annotations contribute towards a comprehensive molecular story about each UniProt entry.

CONCLUSIONS AND FUTURE WORK

Intrinsically disordered regions are key for the function of numerous proteins. High quality experimental disorder annotations can be extracted by manual curation and automatically from the PDB. Due to the difficulties in experimentally characterizing disorder, many computational predictors have been developed with various disorder flavors and are essential for large-scale annotation. Here we provide a new version of MobiDB, a centralized source for data on different flavors of disorder in protein structures now covering over 80 million proteins. The database features three levels of annotation: manually curated, indirect and predicted. The new version also features a consensus annotation for long disordered regions. MobiDB aims at giving the best possible picture of the 'disorder landscape' of a given protein of interest. Since it currently covers the full set of UniProt sequences, the included predictors need to be extremely fast, enabling MobiDB to provide disorder annotations for every protein, especially when no curated or indirect data is available. In order to complement the disorder annotations, MobiDB features additional annotations from external sources like the UniProt, Pfam and STRING databases including domains, protein–protein interactions, post-translational modifications, binding sites and low complexity regions.

Beyond its current release, MobiDB is a continuous effort to expand, revise and improve intrinsically disordered annotations. The maintenance of such an amount of data is not simple, especially if we consider that the number of protein sequences in UniProt has doubled in less than a year, so the main effort will be to maintain a fully automated protocol allowing regular database updates. Inclusion of other prediction types such as amyloid aggregation tendency with PASTA 2.0 (33) or ubiquitinylation with RUBI (34) is also possible. Thematic collections, e.g. proteins for specific organisms and/or annotation types will be provided in due course. Interested users are encouraged to submit requests through the online contact form. MobiDB provides the means to obtain disorder annotations for more than 80 million proteins, providing the highest sequence-coverage of any available database, while annotating intrinsic disorder as well as possible through its combination of experimental sources and consensus predictions.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENT

The authors are grateful to Vladimir Uversky, Giovanni Minervini, Manuel Giollo and the BioComputing Lab for insightful discussions and to A. Keith Dunker for maintaining the DisProt database.

ACCESSION NUMBER

PDB ID: 2kkw.

FUNDING

FIRB Futuro in Ricerca [RBFR08ZSXY to S.T.]; AIRC [MFAG 12740 to S.T.]. Funding for open access charge: FIRB Futuro in Ricerca [RBFR08ZSXY].
Conflict of interest statement. None declared.

REFERENCES

1. Tompa,P. (2012) Intrinsically disordered proteins: a 10-year recap. *Trends Biochem. Sci.*, **37**, 509–516.
2. Habchi,J., Tompa,P., Longhi,S. and Uversky,V.N. (2014) Introducing protein intrinsic disorder. *Chem. Rev.*, **114**, 6561–6588.
3. Ward,J.J., Sodhi,J.S., McGuffin,L.J., Buxton,B.F. and Jones,D.T. (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.*, **337**, 635–645.
4. Dunker,A.K., Obradovic,Z., Romero,P., Garner,E.C. and Brown,C.J. (2000) Intrinsic protein disorder in complete genomes. *Genome Inform. Workshop Genome Inform.*, **11**, 161–171.
5. 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. *et al.* (2014) Classification of intrinsically disordered regions and proteins. *Chem. Rev.*, **114**, 6589–6631.
6. Sickmeier,M., Hamilton,J.A., LeGall,T., Vacic,V., Cortese,M.S., Tantos,A., Szabo,B., Tompa,P., Chen,J., Uversky,V.N. *et al.* (2007) DisProt: the Database of Disordered Proteins. *Nucleic Acids Res.*, **35**, D786–D793.
7. Fukuchi,S., Amemiya,T., Sakamoto,S., Nobe,Y., Hosoda,K., Kado,Y., Murakami,S.D., Koike,R., Hiroaki,H. and Ota,M. (2014) IDEAL in 2014 illustrates interaction networks composed of intrinsically disordered proteins and their binding partners. *Nucleic Acids Res.*, **42**, D320–D325.
8. Rose,P.W., Bi,C., Bluhm,W.F., Christie,C.H., Dimitropoulos,D., Dutta,S., Green,R.K., Goodsell,D.S., Prlic,A., Quesada,M. *et al.* (2013) The RCSB Protein Data Bank: new resources for research and education. *Nucleic Acids Res.*, **41**, D475–D482.
9. Tompa,P. (2002) Intrinsically unstructured proteins. *Trends Biochem. Sci.*, **27**, 527–533.
10. Walsh,I., Giollo,M., Di Domenico,T., Ferrari,C., Zimmermann,O. and Tosatto,S.C.E. (2014) Comprehensive large scale assessment of intrinsic protein disorder. *Bioinformatics*, doi:10.1093/bioinformatics/btu625.
11. 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.
12. Monastyrskyy,B., Kryshtafovych,A., Moult,J., Tramontano,A. and Fidelis,K. (2014) Assessment of protein disorder region predictions in CASP10. *Proteins*, **82**(Suppl. 2), 127–137.
13. Linding,R., Russell,R.B., Neduva,V. and Gibson,T.J. (2003) GlobPlot: exploring protein sequences for globularity and disorder. *Nucleic Acids Res.*, **31**, 3701–3708.
14. Dosztányi,Z., Csizmók,V., Tompa,P. and Simon,I. (2005) The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J. Mol. Biol.*, **347**, 827–839.
15. Walsh,I., Martin,A.J.M., Di Domenico,T. and Tosatto,S.C.E. (2012) ESpritz: accurate and fast prediction of protein disorder. *Bioinformatics*, **28**, 503–509.
16. Yang,Z.R., Thomson,R., McNeil,P. and Esnouf,R.M. (2005) RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics*, **21**, 3369–3376.
17. Linding,R., Jensen,L.J., Diella,F., Bork,P., Gibson,T.J. and Russell,R.B. (2003) Protein disorder prediction: implications for structural proteomics. *Structure*, **11**, 1453–1459.
18. Peng,K., Radivojac,P., Vucetic,S., Dunker,A.K. and Obradovic,Z. (2006) Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics*, **7**, 208.
19. Fukuchi,S., Hosoda,K., Homma,K., Gojobori,T. and Nishikawa,K. (2011) Binary classification of protein molecules into intrinsically disordered and ordered segments. *BMC Struct. Biol.*, **11**, 29.
20. Di Domenico,T., Walsh,I., Martin,A.J.M. and Tosatto,S.C.E. (2012) MobiDB: a comprehensive database of intrinsic protein disorder annotations. *Bioinformatics*, **28**, 2080–2081.
21. Oates,M.E., Romero,P., Ishida,T., Ghalwash,M., Mizianty,M.J., Xue,B., Dosztányi,Z., Uversky,V.N., Obradovic,Z., Kurgan,L. *et al.* (2013) D^2P^2: database of disordered protein predictions. *Nucleic Acids Res.*, **41**, D508–D516.
22. UniProt Consortium (2014) Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **42**, D191–D198.
23. Finn,R.D., Bateman,A., Clements,J., Coggill,P., Eberhardt,R.Y., Eddy,S.R., Heger,A., Hetherington,K., Holm,L., Mistry,J. *et al.* (2014) Pfam: the protein families database. *Nucleic Acids Res.*, **42**, D222–D230.
24. Velankar,S., Dana,J.M., Jacobsen,J., van Ginkel,G., Gane,P.J., Luo,J., Oldfield,T.J., O'Donovan,C., Martin,M.-J. and Kleywegt,G.J. (2013) SIFTS: structure integration with function, taxonomy and sequences resource. *Nucleic Acids Res.*, **41**, D483–D489.
25. Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
26. Wootton,J.C. (1994) Non-globular domains in protein sequences: automated segmentation using complexity measures. *Comput. Chem.*, **18**, 269–285.
27. Jones,D.T. and Swindells,M.B. (2002) Getting the most from PSI-BLAST. *Trends Biochem. Sci.*, **27**, 161–164.
28. Lobley,A., Swindells,M.B., Orengo,C.A. and Jones,D.T. (2007) Inferring function using patterns of native disorder in proteins. *PLoS Comput. Biol.*, **3**, e162.
29. Romero,P., Obradovic,Z., Li,X., Garner,E.C., Brown,C.J. and Dunker,A.K. (2001) Sequence complexity of disordered protein. *Proteins*, **42**, 38–48.
30. Franceschini,A., Szklarczyk,D., Frankild,S., Kuhn,M., Simonovic,M., Roth,A., Lin,J., Minguez,P., Bork,P., von Mering,C. *et al.* (2013) STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.*, **41**, D808–D815.
31. Iakoucheva,L.M., Brown,C.J., Lawson,J.D., Obradović,Z. and Dunker,A.K. (2002) Intrinsic disorder in cell-signaling and cancer-associated proteins. *J. Mol. Biol.*, **323**, 573–584.
32. Di Domenico,T., Potenza,E., Walsh,I., Parra,R.G., Giollo,M., Minervini,G., Piovesan,D., Ihsan,A., Ferrari,C., Kajava,A.V. *et al.* (2014) RepeatsDB: a database of tandem repeat protein structures. *Nucleic Acids Res.*, **42**, D352–D357.
33. Walsh,I., Seno,F., Tosatto,S.C.E. and Trovato,A. (2014) PASTA 2.0: an improved server for protein aggregation prediction. *Nucleic Acids Res.*, **42**, W301–W307.
34. Walsh,I., Di Domenico,T. and Tosatto,S.C.E. (2014) RUBI: rapid proteomic-scale prediction of lysine ubiquitination and factors influencing predictor performance. *Amino Acids*, **46**, 853–862.