Project Acronym:	**IDPfun**
Project Full Title:	**Driving functional characterization of intrinsically disordered proteins**
Grant Agreement:	**778247**
Project Duration:	**48 months (01/03/2018 - 28/02/2022)**

Deliverable D2.2

Software to discriminate different types of disorder based on amino acid propensity and other features

Work Package:	**WP2**
Lead Beneficiary:	**EOTVOS**
Due Date:	**31 May 2020 (M27) - Rev. Due Date 31 Ago 2019 (M30)**
Submission Date:	**5 Ago 2019 (M30)**
Deliverable Type:	**D**
Dissemination Level:	**P**

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

The earlier version of MobiDB-lite is currently used in large-scale proteome annotation platforms to detect intrinsic disorder. However, new theoretical models allow for the classification of intrinsically disordered regions into subtypes from sequence features associated with specific polymeric properties or compositional bias.

MobiDB-lite 3.0 maintains high accuracy and quick execution time but also provides a finer classification of disorder by identifying regions with characteristics of polyolyampholytes, positive or negative polyelectrolytes, low complexity regions or enriched in cysteine, proline or glycine or polar residues. Sub-regions are abundantly detected in IDRs of the human proteome. The new version of MobiDB-lite represents a new step for the proteome level analysis of protein disorder.

Both the MobiDB-lite 3.0 source code and a docker container are available from the GitHub repository: https://github.com/BioComputingUP/MobiDB-lite.

Project overview

The identification of protein domains and sequence conservation has long been central to the annotation of proteomes [1], [2]. Many proteins, known as intrinsically disordered, have been observed to escape the typical organization of globular proteins in domains [3]. A large fraction of the human proteome is devoid of domains [4] and in this 'dark' proteome molecular conformations are completely unknown [5]. Computational prediction of intrinsic disorder (ID) attempts to fill this gap by offering a wide array of prediction methods [6], [7] with different performances [8], [9]. Despite many methods having been available for a long time, they had not been integrated into large-scale proteome annotation [10] was the first of such predictors to be included in InterProScan from its release 60 [11]. MobiDB-lite combines a set of complementary ID predictors in a consensus optimized on a PDB X-ray dataset [8] to limit over-prediction while balancing under-prediction.
In recent years theoretical models of ID proteins surpassed the bare distinction between disorder and structure and reached a point where classification of subtypes of disorder is possible based on sequence features [12], [13]. Furthermore, recent evidence highlighted how ID and low sequence complexity (LC) are strictly intertwined [14]. For this reason, we developed a new version of MobiDB-lite, which can capture different classes of disorder and sequence features that we observed being biologically relevant in IDPs [15]. MobiDB-lite 3.0 is already included in the latest versions of MobiDB [16]. The new version is available as docker container and also exposes bindings to use MobiDB-lite as a python library, in compliance with the FAIR principles [17].

Implementation

MobiDB-lite disorder prediction unfolds in two steps as explained as in [10]. Briefly, the first step calculates a consensus between 8 predictors, which in the second step is smoothed out and filtered to keep only regions longer than 20 residues. MobiDB-lite 3.0 further processes predicted disordered regions in order to achieve a finer classification in sub-regions of at least 5 residues. Each amino acid in the sequence of disordered regions is assigned to one of six classes based on conditions tested on a sliding window of 15 residues. A residue can only be assigned to a single class even if conditions for multiple classes are satisfied. Classes, in order of priority are: Polyampholyte, Positive Polyelectrolyte, Negative Polyelectrolyte, Cysteine-rich, Proline-rich, Glycine-rich, Low complexity, Polar. The first three classes reflect a classification proposed in [12] and were suggested to be associated with different structural and potentially functional characteristics. The latter four are assigned when the fraction of cysteines, prolines, glycines and polar residues in the sliding window is greater than 32%. Finally, Low-complexity class is predicted by SEG [18]. Both the dimension of the sliding window and the threshold on the fraction of residues in the sliding window were manually set based on a sample of biologically relevant proteins. Classification (including SEG prediction) is smoothed out in an iterative process following the same approach applied to disordered regions [10]. Finally, a sequence filter is applied to select only regions longer than 4 residues. Both the MobiDB/lite source code and a docker container are available in the GitHub repository: https://github.com/BioComputingUP/MobiDB-lite.

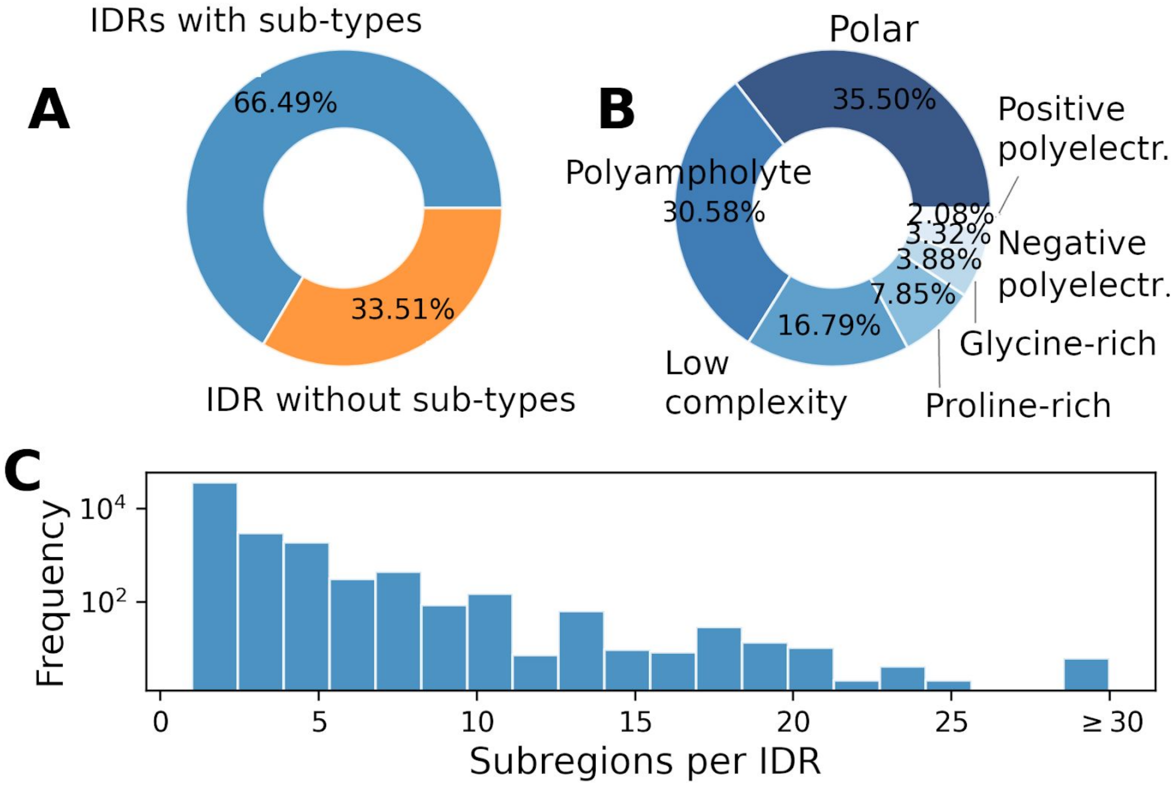

Fig. 1. Abundance of IDRs and sub-regions. MobiDB-lite results for the human proteome. (**A**) Fraction of IDRs with (blue) and without sub-types (orange). (**B**) Distribution of IDR sub-types . (**C**) Distribution of sub-regions detected per IDR, plotted on a logarithmic frequency scale (y-axis).

Use case

IDRs and sub-regions were calculated with MobiDB-lite 3.0 for the whole human proteome from UniProt (UP000005640), consisting of 74,043 amino acid sequences, of which 33,322 (44.5%) are predicted IDPs. A total of 64,484 IDRs and 71,921 sub-regions were detected. The majority of IDPs (59.9%) have just one IDR while only 6 proteins have more than 30 IDRs. Mucin-16 (14,451 residues; UniProt ID: Q8WXI7) contains 112 IDRs. Of the 64,484 IDRs detected, 21,610 (33.5%) do not have any sub-regions, while the remaining 66.5% can have 1 or more (Figure 1A). More than 66.1% of sub-regions are either Polyampholytes or Polar (Figure 1B). The remaining sub-regions are, in order of abundance, low-complexity (16.8%), proline-rich (7.9%), glycine-rich (3.9%), negative (3.3%) and positive polyelectrolytes (2.1%). Cysteine-rich sub-regions are never detected in this dataset. Many IDRs (67.6%) have just one sub-region and the number of IDRs with more than one sub-region drops exponentially with the increase of sub-regions (Figure 1C). In only six cases an IDR has more than 30 sub-regions. Filaggrin (4,061 residues; UniProt ID: P20930) has a predicted IDR spanning from residue 255 to 3,971 hosting 105 sub-regions.

Conclusions

MobiDB-lite 3.0 is a highly specific and very fast consensus predictor achieving a finer ID classification by detecting sub-regions in predicted IDRs. To the best of our knowledge, MobiDB-lite 3.0 is currently the only ID predictor able to sub-classify disorder and also the first ID predictor provided as a docker container.

References

[1] D. Lee, A. Grant, R. L. Marsden, and C. Orengo, “Identification and distribution of protein families in 120 completed genomes using Gene3D,” *Proteins*, vol. 59, no. 3, pp. 603–615, May 2005, doi: 10.1002/prot.20409.

[2] E. L. Sonnhammer, S. R. Eddy, and R. Durbin, “Pfam: a comprehensive database of protein domain families based on seed alignments,” *Proteins*, vol. 28, no. 3, pp. 405–420, Jul. 1997.

[3] A. K. Dunker *et al.*, “Intrinsically disordered protein,” *J. Mol. Graph. Model.*, vol. 19, no. 1, pp. 26–59, Feb. 2001, doi: 10.1016/S1093-3263(00)00138-8.

[4] J. Mistry *et al.*, “The challenge of increasing Pfam coverage of the human proteome,” *Database*, vol. 2013, no. 0, pp. bat023–bat023, Apr. 2013, doi: 10.1093/database/bat023.

[5] N. Perdigão *et al.*, “Unexpected features of the dark proteome,” *Proc. Natl. Acad. Sci.*, p. 201508380, Nov. 2015, doi: 10.1073/pnas.1508380112.

[6] I. Walsh, A. J. M. Martin, T. Di Domenico, and S. C. E. Tosatto, “ESpritz: accurate and fast prediction of protein disorder,” *Bioinformatics*, vol. 28, no. 4, pp. 503–509, Feb. 2012, doi: 10.1093/bioinformatics/btr682.

[7] B. Mészáros, G. Erdős, and Z. Dosztányi, “IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding,” *Nucleic Acids Res.*, vol. 46, no. W1, pp. W329–W337, Jul. 2018, doi: 10.1093/nar/gky384.

[8] I. Walsh, M. Giollo, T. Di Domenico, C. Ferrari, O. Zimmermann, and S. C. E. Tosatto, “Comprehensive large-scale assessment of intrinsic protein disorder,” *Bioinformatics*, vol. 31, no. 2, pp. 201–208, Jan. 2015, doi: 10.1093/bioinformatics/btu625.

[9] M. Necci, D. Piovesan, Z. Dosztányi, P. Tompa, and S. C. E. Tosatto, “A comprehensive assessment of long intrinsic protein disorder from the DisProt database,” *Bioinformatics*, vol. 34, no. 3, pp. 445–452, 01 2018, doi: 10.1093/bioinformatics/btx590.

[10] M. Necci, D. Piovesan, Z. Dosztányi, and S. C. E. Tosatto, “MobiDB-lite: fast and highly specific consensus prediction of intrinsic disorder in proteins,” *Bioinformatics*, vol. 33, no. 9, pp. 1402–1404, May 2017, doi: 10.1093/bioinformatics/btx015.

[11] A. L. Mitchell *et al.*, “InterPro in 2019: improving coverage, classification and access to protein sequence annotations,” *Nucleic Acids Res.*, vol. 47, no. D1, pp. D351–D360, 08 2019, doi: 10.1093/nar/gky1100.

[12] R. K. Das and R. V. Pappu, “Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 110, no. 33, pp. 13392–13397, Aug. 2013, doi: 10.1073/pnas.1304749110.

[13] A. S. Holehouse, R. K. Das, J. N. Ahad, M. O. G. Richardson, and R. V. Pappu, “CIDER: Resources to Analyze Sequence-Ensemble Relationships of Intrinsically

Disordered Proteins," *Biophys. J.*, vol. 112, no. 1, pp. 16–21, Jan. 2017, doi: 10.1016/j.bpj.2016.11.3200.
[14] P. Mier *et al.*, "Disentangling the complexity of low complexity proteins," *Brief. Bioinform.*, vol. 21, no. 2, pp. 458–472, Mar. 2020, doi: 10.1093/bib/bbz007.
[15] M. Necci, D. Piovesan, and S. C. E. Tosatto, "Large-scale analysis of intrinsic disorder flavors and associated functions in the protein sequence universe," *Protein Sci. Publ. Protein Soc.*, vol. 25, no. 12, pp. 2164–2174, 2016, doi: 10.1002/pro.3041.
[16] D. Piovesan *et al.*, "MobiDB 3.0: more annotations for intrinsic disorder, conformational diversity and interactions in proteins," *Nucleic Acids Res.*, vol. 46, no. D1, pp. D471–D476, Jan. 2018, doi: 10.1093/nar/gkx1071.
[17] M. D. Wilkinson *et al.*, "The FAIR Guiding Principles for scientific data management and stewardship," *Sci. Data*, vol. 3, p. 160018, Mar. 2016, doi: 10.1038/sdata.2016.18.
[18] J. C. Wootton and S. Federhen, "Statistics of local complexity in amino acid sequences and sequence databases," *Comput. Chem.*, vol. 17, no. 2, pp. 149–163, Jun. 1993, doi: 10.1016/0097-8485(93)85006-X.