bioRxiv preprint doi: https://doi.org/10.1101/060699; this version posted March 26, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Disorder Atlas: web-based software for the proteome-based interpretation of intrinsic disorder predictions

Michael Vincent[1] & Santiago Schnell[1,2,3]

[1] Department of Molecular & Integrative Physiology, University of Michigan Medical School, Ann Arbor, MI, USA
[2] Department of Computational Medicine & Bioinformatics, University of Michigan Medical School, MI, USA
[3] Brehm Center for Diabetes Research, University of Michigan Medical School, Ann Arbor, MI, USA

To whom correspondence should be addressed: Santiago Schnell, Brehm Center 5132, 1000 Wall Street, Ann Arbor, Michigan 48105-1912, USA. Telephone: (734) 615-8733; Fax: (734) 232-8162; Email: schnells@umich.edu

---

Abstract

Background:
Intrinsically disordered proteins lack a stable three-dimensional structure under physiological conditions. While this property has gained considerable interest within the past two decades, disorder poses substantial challenges to experimental characterization efforts. In effect, numerous computational tools have been developed to predict disorder from primary sequences, however, interpreting the output of these algorithms remains a challenge.

Results:
Here, we present Disorder Atlas, web-based software that facilitates the interpretation of intrinsic disorder predictions using proteome-based descriptive statistics. This service is also equipped to facilitate large-scale systematic exploratory searches for proteins encompassing disorder features of interest, and further allows users to browse the prevalence of multiple disorder features at the proteome level. Disorder Atlas is freely available for non-commercial users at http://www.disorderatlas.org.

Conclusion:
Disorder Atlas enables a standardized interpretation of intrinsic disorder predictions, and provides researchers with a tool to assess disorder on multiple scales.

Keywords: intrinsic disorder, protein disorder, standardization, disorder prediction, proteome, continuous disorder, IUPred, DisEMBL, consensus disorder.

Background

Intrinsically disordered proteins and protein regions do not form a stable three-dimensional structure under physiological conditions. A number of prediction algorithms capable

bioRxiv preprint doi: https://doi.org/10.1101/060699; this version posted March 26, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

of predicting disorder in protein sequences play a critical role in disorder characterization efforts [1, 2]. When analyzing the results users are faced with the problem of interpreting the predicted disorder content of a protein, which often involves assessing the percent disorder, and length and location of continuous stretches of disordered residues. However, if 'protein X' is predicted to contain a disorder content of 20%, how does one evaluate whether this feature is significant?

Disorder Atlas is web-based software that facilitates the interpretation of disorder predictions from amino acid sequence by comparing them with descriptive statistics, specific to both proteomes and disorder prediction tools, for identifying anomalous disorder features with respect to whole proteomic populations. We adopt the noun "atlas" under Gerardus Mercator's original neologism of a descriptive reference rather than an extensive or complete collection. In Disorder Atlas, the reference is a set of proteome-based guidelines for standardizing intrinsic disorder predictions [3], which are analogous to clinical guidelines used to evaluate whether an individual is overweight based on the body mass index distribution in the population. Although these guidelines do not provide a functional role of the disorder predictions, they provide a standardization of the prevalence of disorder in a protein with respect to its proteome. These guidelines align with the intent of the US National Institutes of Health's "Principles and Guidelines for Reporting Preclinical Research" to encourage the use of standards for rigorous reporting of research results [4].

Implementation

Disorder Atlas is primarily written in Python 2.7.12 and utilizes the Django model-view-controller framework. Disorder Atlas also interfaces with algorithms written in C and graphics modules written in JavaScript. All database operations are carried out using the Structured Query Language (SQL).

Pre-calculated proteome statistics are pulled from a PostgreSQL database, which have been computed from ten representative eukaryotic protein populations with minimal sequence redundancy and uncertainty [3, 5]. Specifically, Disorder Atlas includes the *Saccharomyces cerevisiae*, *Dictyostelium discoideum*, *Chlamydonmonas reinhardtii*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Arabidopsis thaliana*, *Danio rerio*, *Mus musculus*, *Homo sapiens*, and *Zea mays* proteomes. However, Disorder Atlas accepts input sequences from any organism.

Reference proteome files obtained from UniProt served as the source of all protein sequence information and were subjected to multiple levels of filtering prior to disorder analysis. Disorder Atlas defines a sequence to be eligible for analysis if it does not include ambiguous and uncertain amino acid residues (such as B, J, O, U, X, and Z), and excludes sequences containing these features due to the variability observed in disorder prediction algorithms in processing input sequences containing these residue types [3, 5]. The rationale behind this exclusion is that algorithmic variability for ambiguous and uncertain amino acids introduces error jeopardizing the accuracy of the analysis and standardization. Furthermore, redundancy was minimized by using UniRef100 reference cluster information obtained from the UniProt identification mapping service [6, 7]. The redundancy minimization procedure has been described in detail elsewhere [3].

Currently, Disorder Atlas supports IUPred [8, 9] and DisEMBL [10] disorder predictions (as well as consensus agreement between the two). Briefly, IUPred predicts disorder by assessing pairwise interresidue interaction energy and identifies disordered residues as those that are incapable of forming stabilizing interresidue interactions [8, 9]. DisEMBL is a suite of three

bioRxiv preprint doi: https://doi.org/10.1101/060699; this version posted March 26, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

neural networks algorithms that predict the presence of disorder: COILS, HOTLOOPS, and REM465, referred to by Disorder Atlas as DisEMBL-C, DisEMBL-H, and DisEMBL-R, respectively. Each describes disorder as a two-state model, classifying residues as either disordered or ordered [10]. DisEMBL-C utilizes secondary structure prediction to assign disorder/order classifications, and classifies residues as disordered if they are present in loops [10]. Residues are predicted to belong to loops if they do not belong to either alpha-helices, 3_{10}-helices, or beta-strands (assumes all residues belonging to loops are disordered) [10]. Building on the reasoning that not all loops are disordered but all disordered residues are found within loops, DisEMBL-C classifications represent an overestimate of disorder. For improved disorder classifications, Linding et al. [10], implemented DisEMBL- H, which classifies residues (contained within loops) as disordered only if its alpha-carbon has a high temperature factor (B factor). Lastly, the DisEMBL-R neural network has been trained using non-assigned electron densities from X-ray crystallography (XRC) data contained within the Protein Data Bank, and assumes residues with missing XRC coordinates (as defined by REMARK 465 entries in PDB files) are disordered [10].

Disorder Atlas also provides information about protein hydropathy, charge distribution and other disorder-relevant parameters as predicted by CIDER [11]. In addition to its single protein assessment features, Disorder Atlas also provides tools for browsing disorder at the proteome-level and for conducting an exploratory search for proteins with disorder features of interest. These tools are described in detail in the next section.

Results and Discussion

The Disorder Atlas web-based interface provides access to three tools for interpreting disorder predictions: (1) a proteome-level disorder browser, (2) an individual protein analysis tool, and (3) a proteome exploratory search tool. Each tool is based on descriptive population statistics, and presents the standing of disorder features in relation to a proteomic population based on quantitative guidelines for standardizing disorder predictions [3]. Disorder Atlas utilizes two physicochemical-based disorder prediction algorithms, IUPred-L and DisEMBL. Consensus disorder predictions are also presented, which provide more conservative disorder annotations.

The Proteome Browser

This tool provides the distribution of three disorder statistics at the proteome level: (1) the disorder content, (2) the longest continuous disorder region (CD_L), and (3) the longest CD_L percentage of length (LCPL). The disorder content is simply the percentage of disordered residues contained within a protein sequence. The CD_L is the longest continuously disordered region in a protein, defined using the theoretical minimum of two consecutive disordered residues (while a CD segment of two amino acids may be structurally unimportant, it includes all possible predicted CD segments and avoids using a subjective minimum length that could potentially exclude valid short CD regions). The LCPL defines the percentage of the total protein length accounted for by the CD_L and is useful for identifying a statistically relevant long CD segment in proteins having a primary sequence length exceeding previously reported protein length thresholds [3]. For each of these disorder statistics, users can access the proteome browser to visualize statistical distributions, percentiles, and expected values.

bioRxiv preprint doi: https://doi.org/10.1101/060699; this version posted March 26, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Individual Protein Analysis Tool

For single protein analyses, users can provide either the (1) UniProt accession number, or (2) FASTA sequence, and the name of the proteome to which the sequence belongs. Following submission, the disorder propensity, as well as the standing of the disorder content, CD_L, and LCPL with respect to the proteome, is presented (two sample result plots are displayed in **FIGURE 1A**). Additionally, Disorder Atlas also presents protein hydropathy and charge distribution and other disorder-relevant parameters generated by localCIDER version 0.1.7 [11]. The generated histograms, boxplots, and disorder propensity charts can be downloaded as either a PNG or SVG file. All pages can be easily saved as a PDF file from any web browser print menu.

Proteome Exploratory Search Tool

Disorder Atlas can provide an exploratory search for proteins with a disorder feature of interest. To conduct this search, users specify the proteome to be searched, the disorder metric and prediction method of interest, and whether they would like to conduct a value-based or percentile-based search. For example, a user could search for all *Saccharomyces cerevisiae* proteins with a DisEMBL-R-predicted percent disorder of less than 40% (a truncated result table for this search is displayed in **FIGURE 1B**), or they could look for all *Homo sapiens* proteins with a CD_L above the 75th percentile (not shown). After submitting a search query Disorder Atlas returns a list of proteins meeting the specified criteria together with their associated statistical values. The search results can be exported in a variety of file formats, including CSV, JSON, PDF, SQL, TXT, and XML.

Improvements over existing protein disorder resources

A handful of databases exist that contain disorder annotations, including D2P2 [12], DisProt [13], and MobiDB [14, 15]. These resources provide users with valuable information concerning protein regions predicted to exhibit disorder, and in some cases, provide experimental support of disorder predictions. Importantly, the number of protein sequences having experimental information relevant to protein disorder represent only a small fraction of the total number of known sequences. As such, researchers are often confronted with the problem of interpreting computational disorder predictions without this information.

Intrinsic disorder is a prevalent feature of many proteomes [3, 5], and such, the mere prediction of disorder (to any degree) is insignificant. In cases where experimental information regarding the disorder content for a protein of interest is unavailable, software is needed to enable a standardized interpretation of these predictions. For example, if a given protein is found to contain 20 amino acid long disordered region, is this significant at the proteome-level? Or if 30% of the residues in a protein are predicted to be disordered, what should be made of this prediction from the proteome perspective? The aforementioned databases cannot answer these questions when experimental information is unavailable. Disorder Atlas addresses a major need that is unmet by any existing resources containing disorder annotations: it provides researchers with a tool for standardizing disorder predictions with respect to whole proteomes.

Conclusions

Numerous intrinsic disorder prediction algorithms exist and as our understanding of disorder expands, more algorithms are developed. Yet, limited guidelines for standardizing the

bioRxiv preprint doi: https://doi.org/10.1101/060699; this version posted March 26, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

prevalence of disorder restrict the interpretation of predictions by scientists without a sophisticated understanding of structural biology and protein informatics. Disorder Atlas is web-based software that aims to bridge this gap by providing accessible and versatile tools for standardizing protein disorder predictions. The guidelines associated to Disorder Atlas provide a meaningful improvement in the methods of reporting protein disorder. They could increase the value of intrinsically disordered predictions in accordance to transparent preclinical reporting recommendations [16].

We plan to support additional proteomes within the upcoming year after implementing an automated disorder prediction and analysis pipeline. Following implementation of this system, users will have access to a multitude of prokaryotic and eukaryotic proteomes. We further envision that additional disorder prediction algorithms will be supported in the future as well.

Declarations

Funding
This work was partially supported by the University of Michigan Protein Folding Diseases Initiative and the Research Discovery Fund.

Ethics approval and consent to participate
Not applicable

Consent for publication
Not applicable

Availability of data and material
The protein sequences, their disorder predictions, hydropathy, charge distribution and other disorder-relevant parameters that support Disorder Atlas are available in Dryad with the identifier DOI 10.5061/dryad.sm107. [5]

Disorder Atlas is freely available for non-commercial users at http://www.disorderatlas.org. The computer code is published with a dual license scheme (GNU General Public License and a modified version of the Limited GNU Public License) to permit use and further development by non-profit and commercial entities.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
MV designed, developed, and produced Disorder Atlas. SS contributed to the design and supervised the project. MV and SS wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements
We thank Dr. Arash S. Soleimanpour (University of Michigan) for testing Disorder Atlas.

bioRxiv preprint doi: https://doi.org/10.1101/060699; this version posted March 26, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

References

1. Monastyrskyy B, Kryshtafovych A, Moult J, Tramontano A, Fidelis K: **Assessment of protein disorder region predictions in CASP10**. *Proteins* 2014, **82 Suppl 2**:127-137.
2. Atkins JD, Boateng SY, Sorensen T, McGuffin LJ: **Disorder Prediction Methods, Their Applicability to Different Protein Targets and Their Usefulness for Guiding Experimental Studies**. *Int J Mol Sci* 2015, **16**(8):19040-19054.
3. Vincent M, Whidden M, Schnell S: **Quantitative proteome-based guidelines for intrinsic disorder characterization**. *Biophys Chem* 2016, **213**:6-16.
4. **Principles and Guidelines for Reporting Preclinical Research** [https://www.nih.gov/research-training/rigor-reproducibility/principles-guidelines-reporting-preclinical-research]
5. Vincent M, Schnell S: **A collection of intrinsic disorder characterizations from eukaryotic proteomes**. *Sci Data* 2016, **3**:160045.
6. Suzek B, Huang H, McGarvey P, Mazumder R, Wu C: **UniRef: comprehensive and non-redundant UniProt reference clusters.** *Bioinformatics* 2007, **23**(10):1282-1288.
7. UniProt C: **UniProt: a hub for protein information**. *Nucleic Acids Res* 2015, **43**(Database issue):D204-212.
8. Dosztanyi Z, Csizmok V, Tompa P, Simon I: **The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins**. *J Mol Biol* 2005, **347**(4):827-839.
9. Dosztanyi Z, Csizmok V, Tompa P, Simon I: **IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content**. *Bioinformatics* 2005, **21**(16):3433-3434.
10. Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB: **Protein disorder prediction: implications for structural proteomics**. *Structure* 2003, **11**(11):1453-1459.
11. Holehouse AS, Das RK, Ahad J, Richardson MOG, Pappu RV: **CIDER: Resources to analyze sequence-ensemble relationships of intrinsically disordered proteins**. *Biophysical Journal* 2017, **112**:16-21.
12. Oates ME, Romero P, Ishida T, Ghalwash M, Mizianty MJ, Xue B, Dosztanyi Z, Uversky VN, Obradovic Z, Kurgan L *et al*: **D(2)P(2): database of disordered protein predictions**. *Nucleic Acids Res* 2013, **41**(Database issue):D508-516.
13. Piovesan D, Tabaro F, Micetic I, Necci M, Quaglia F, Oldfield CJ, Aspromonte MC, Davey NE, Davidovic R, Dosztanyi Z *et al*: **DisProt 7.0: a major update of the database of disordered proteins**. *Nucleic Acids Res* 2017, **45**(D1):D219-D227.
14. Di Domenico T, Walsh I, Martin AJ, Tosatto SC: **MobiDB: a comprehensive database of intrinsic protein disorder annotations**. *Bioinformatics* 2012, **28**(15):2080-2081.
15. Potenza E, Di Domenico T, Walsh I, Tosatto SC: **MobiDB 2.0: an improved database of intrinsically disordered and mobile proteins**. *Nucleic Acids Res* 2015, **43**(Database issue):D315-320.

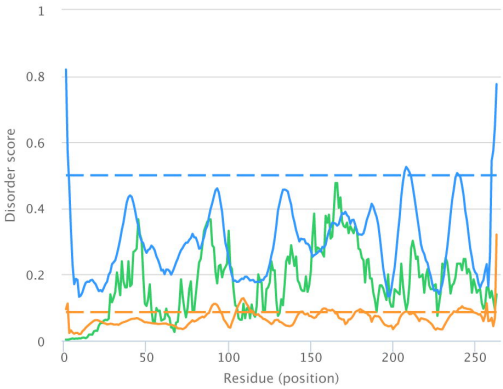
bioRxiv preprint doi: https://doi.org/10.1101/060699; this version posted March 26, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

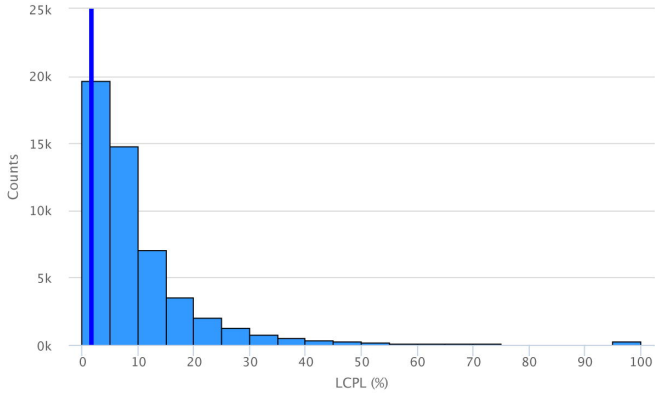
16. Landis SC, Amara SG, Asadullah K, Austin CP, Blumenstein R, Bradley EW, Crystal RG, Darnell RB, Ferrante RJ, Fillit H *et al*: **A call for transparent reporting to optimize the predictive value of preclinical research**. *Nature* 2012, **490**(7419):187-191.

Figure caption

Figure 1. (A) Example output from the individual protein analysis tool. The disorder propensity predicted by IUPred-L (green), DisEMBL-H (orange), and DisEMBL-R (blue) for Chymotrypsinogen B (P17538) is shown on the left, whereas its LCPL predicted by DisEMBL-R (blue vertical line) is shown together with the *Homo sapiens* DisEMBL-R LCPL distribution on the right. **(B)** Example output from the exploratory search tool. A search was conducted to find *Saccharomyces cerevisiae* proteins with a DisEMBL-R-predicted disorder percentage of less than 40%. A truncated table displaying the six most disordered proteins meeting the search criteria is shown.

A





B

UniProtKB Accessio...	Gene name	Protein name	Value	Percentile	PE	SV
P40325	HUA1	Proline-rich protein HUA1	39.9	96	1	2
P39943	MIG3	Transcription corepressor MIG3	39.85	96	1	1
P38824	COX23	Cytochrome c oxidase-assembly factor COX23, mitochond...	39.74	96	1	1
Q2V2P0	YPR145C-A	Uncharacterized protein YPR145C-A	39.74	96	4	1
P32499	NUP2	Nucleoporin NUP2	39.72	96	1	2
P47001	CIS3	Cell wall mannoprotein CIS3	39.65	96	1	1