Disorder Atlas: a web service 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

Summary: Disorder Atlas is a web-based service 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.

1. Introduction

Intrinsically disordered proteins and protein regions do not form a stable three-dimensional structure under physiological conditions. A number of prediction algorithms capable of predicting disorder in protein sequences play a critical role in disorder characterization efforts (Atkins, et al., 2015; Monastyrskyy, et al., 2014). 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 a web-based service 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 developed proteome-based guidelines to interpret intrinsic disorder predictions (Vincent, et al., 2016), 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 help to understand the prevalence of disorder in a protein with respect to its proteome. Currently, Disorder Atlas supports IUPred (Dosztanyi, et al., 2005; Dosztanyi, et al., 2005) and DisEMBL (Linding, et al., 2003) disorder predictions (as well as consensus agreement between the two), and also offers information about protein hydropathy, charge distribution and other disorder-relevant parameters as predicted by CIDER (Holehouse, et al., 2015). 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.

2. Usage and implementation

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 interpreting disorder predictions (Vincent, et al., 2016). 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.

2.1. 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 (Vincent, et al., 2016). For each of these disorder statistics, users can access the proteome browser to visualize statistical distributions, percentiles, and expected values. Disorder Atlas pulls proteome statistics from a PostgreSQL database, which have been calculated from protein populations with minimal sequence redundancy and uncertainty (Vincent and Schnell, 2016; Vincent, et al., 2016).

2.2. 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 (Holehouse, et al., 2015). 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.

2.3. 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.

3. Conclusions and future direction

Numerous intrinsic disorder prediction algorithms exist and as our understanding of disorder expands, more algorithms are developed. Yet, limited guidelines for understanding the prevalence of disorder restrict the interpretation of predictions by scientists without a sophisticated understanding of structural biology and protein informatics. Disorder Atlas is a web service that aims to bridge this gap by providing accessible and versatile tools for interpreting protein disorder predictions.

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.

Funding

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

Acknowledgements

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

References

Atkins, J.D., *et al.* 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.

Dosztanyi, Z., *et al.* IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* 2005;21(16):3433-3434.

Dosztanyi, Z., *et al.* The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J Mol Biol* 2005;347(4):827-839.

Holehouse, A.S., *et al.* CIDER: Classification of Intrinsically Disordered Ensemble Regions. 2015.

Linding, R., *et al.* Protein disorder prediction: implications for structural proteomics. *Structure* 2003;11(11):1453-1459.

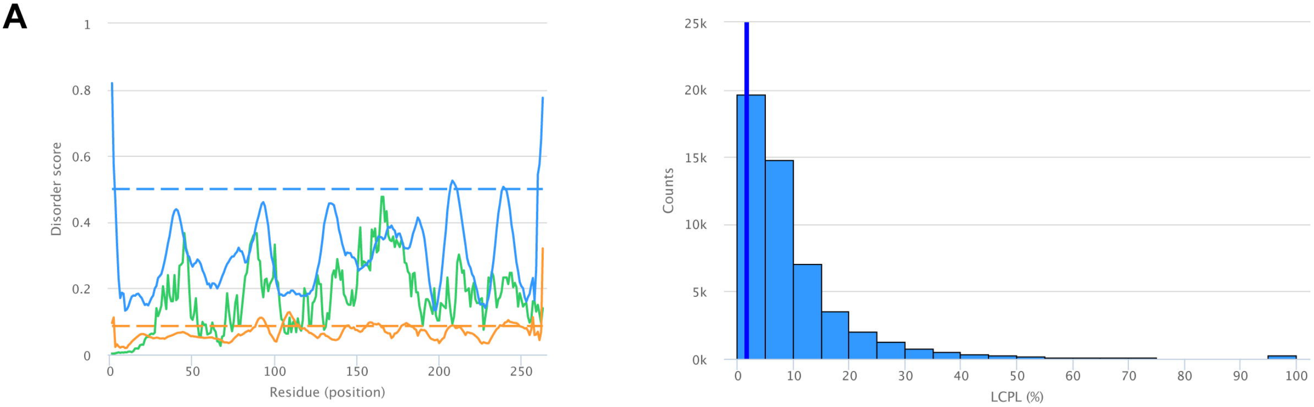
Monastyrskyy, B., *et al.* Assessment of protein disorder region predictions in CASP10. *Proteins* 2014;82 Suppl 2:127-137.

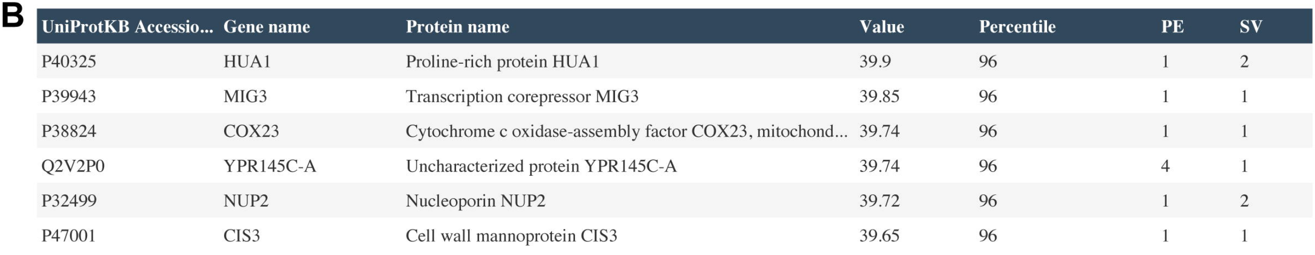
Vincent, M. and Schnell, S. A collection of intrinsic disorder characterizations from eukaryotic proteomes. *Sci Data* 2016;3:160045.

Vincent, M., Whidden, M. and Schnell, S. Quantitative proteome-based guidelines for intrinsic disorder characterization. *Biophys Chem* 2016;213:6-16.

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.





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