Biophysical Chemistry 213 (2016) 6–16

Contents lists available at ScienceDirect

Biophysical Chemistry

journal homepage: http://www.elsevier.com/locate/biophyschem

# Quantitative proteome-based guidelines for intrinsic disorder characterization

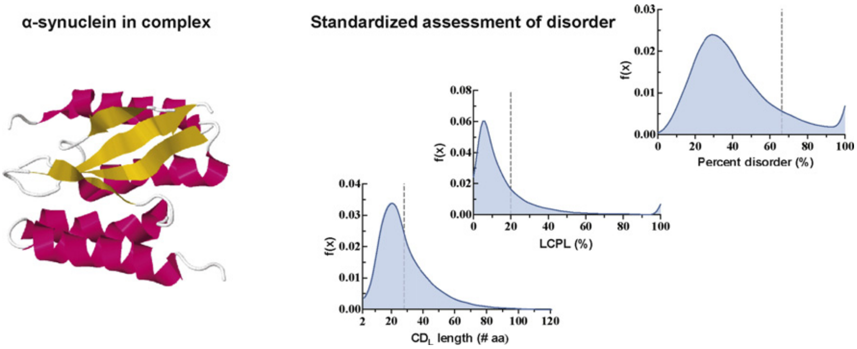
Michael Vincent [a], Mark Whidden [a], Santiago Schnell [a,b,c,*]

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

## HIGHLIGHTS

- A rigorous nonparametric statistical analysis of intrinsic disorder is presented.
- Disorder content and continuous disorder are analyzed in ten eukaryotic proteomes.
- Quantitative guidelines are established for characterizing intrinsic disorder.
- Algorithm-specific expected values and percentile cutoffs are explicitly provided.

## GRAPHICAL ABSTRACT



## ARTICLE INFO

*Article history:*
Received 11 February 2016
Received in revised form 8 March 2016
Accepted 29 March 2016
Available online 5 April 2016

*Keywords:*
Bioinformatics
Computational biology
Intrinsically disordered protein
Protein sequence
Protein structure
Proteomics

## ABSTRACT

Intrinsically disordered proteins fail to adopt a stable three-dimensional structure under physiological conditions. It is now understood that many disordered proteins are not dysfunctional, but instead engage in numerous cellular processes, including signaling and regulation. Disorder characterization from amino acid sequence relies on computational disorder prediction algorithms. While numerous large-scale investigations of disorder have been performed using these algorithms, and have offered valuable insight regarding the prevalence of protein disorder in many organisms, critical proteome-based descriptive statistical guidelines that would enable the objective assessment of intrinsic disorder in a protein of interest remain to be established. Here we present a quantitative characterization of numerous disorder features using a rigorous non-parametric statistical approach, providing expected values and percentile cutoffs for each feature in ten eukaryotic proteomes. Our estimates utilize multiple ab initio disorder prediction algorithms grounded on physicochemical principles. Furthermore, we present novel threshold values, specific to both the prediction algorithms and the proteomes, defining the longest primary sequence length in which the significance of a continuous disordered region can be evaluated on the basis of length alone. The guidelines presented here are intended to improve the interpretation of disorder content and continuous disorder predictions from the proteomic point of view.

## 1. Introduction

Once translated, many nascent unfolded polypeptides fold into a highly ordered conformation. However, within the last two decades it

* Corresponding author at: Brehm Center 5132, 1000 Wall Street, Ann Arbor, MI 48105-1912, USA.
*E-mail address:* schnells@umich.edu (S. Schnell).

http://dx.doi.org/10.1016/j.bpc.2016.03.005
0301-4622/© 2016 Elsevier B.V. All rights reserved.

has been become increasingly apparent that not all proteins fold into a stable globular structure [1–3]. Rather, many proteins and/or protein regions are thought exhibit intrinsic disorder. Intrinsically disordered proteins or protein regions are those that lack a stable three-dimensional structure under physiological conditions, but instead, exist in a natively unfolded state. From a physicochemical standpoint, disordered regions are often characterized by low complexity and the absence of secondary structure, and often consist of residues with low hydrophobicity and high polarity and charge [4]. Disorder has emerged as a prevalent and important feature in the proteomes of many prokaryotes and eukaryotes. Regarding the latter, it has been estimated that 15–45% of eukaryotic proteins contain "significant" long disordered regions, commonly defined as a disordered stretch of 30 or more amino acids in length [5].

While writing off intrinsically disordered proteins as lacking function would be easy due to the absence of a well-defined tertiary structure, a growing body of evidence supports intrinsically disordered proteins playing important functional roles in various signaling and regulatory processes [4,6,7], including apoptosis [8,9], and cell cycle regulation [10]. Interestingly, disorder may also serve as a recognizable feature. Ube2W, a unique ubiquitin-conjugating enzyme (E2) that mono-ubiquitinates the amino-terminus of target substrates, was recently found to specifically recognize substrates with disordered *N*-termini in vitro [11]. Additional support has been established in vivo in a Ube2W knockout mouse model, where both full-length and *N*-terminal disorder were found to be more prevalent in a subset of testicular proteins exhibiting a 1.5× expression increase in the knock-out compared to wild-type [12]. Some proteins involved in protein misfolding diseases are now understood as being intrinsically disordered as well, including the Amyloid-β peptide in Alzheimer's disease and α-synuclein in Parkinson's disease [13].

While analyzing the role of disorder within a single protein or a small set of related proteins is important for understanding the contributions of disorder to protein structure (or the lack of structure) and function, studies must be carried out at the proteomic level to establish critical reference points for disorder characterization. Indeed, proteomic investigations of disorder have been performed and have offered valuable insight into the prevalence of disorder in many organisms [14–16]. However, these studies have not provided guidelines in the form of explicit descriptive statistics, specific to both proteomes and disorder prediction tools, for identifying anomalous disorder features with respect to whole proteomic populations. Without these guidelines in hand, it remains very difficult to understand whether or not a given disorder measure is significant with respect to the population. Guidelines of this nature would be analogous to clinical guidelines used to identify and evaluate whether an individual is overweight or obese based on the body mass index distribution in the population [17–19]. For example, if a protein of interest is found to contain a disordered region that is 25 amino acids in length, is this significant? And how does the context of the primary sequence length influence the evaluation of significance? Before these questions can be answered objectively, a rigorous descriptive statistical analysis of disorder content and continuous disorder must be conducted at the proteome level.

Motivated by these considerations, we analyzed disorder in the proteomes of ten eukaryotic model organisms using a non-parametric descriptive statistical approach. Disorder was estimated using two reputable disorder prediction algorithms, IUPred and DisEMBL, which have a physicochemical basis. While larger-scale disorder studies have been performed, limiting our study to a manageable number of common eukaryotes allowed us to ascertain the quality of the protein sequence pool, quantitatively and qualitatively inspect the accuracy of our statistical methodology, and present objective guidelines for disorder classification in an explicit fashion. This work provides one of the most systematic non-parametric efforts toward standardizing disorder content and continuous length disorder that has been described in the literature.

## 2. Materials and methods

### 2.1. Proteomes and protein sequences

Primary sequences for all proteins included in our analysis were obtained from UniProt reference proteome files [20]. The ability to visualize data distributions in our study is extremely important for testing and presenting the validity of our nonparametric statistical approach, thereby limiting our study to the proteomes of ten model eukaryotes. Specifically, the *Saccharomyces cerevisiae*, *Dictyostelium discoideum*, *Chlamydonmonas reinhardtii*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Arabidopsis thaliana*, *Danio rerio*, *Mus musculus*, *Homo sapiens*, *and Zea mays* proteomes were included in our investigation (proteome presentation order was decided by final protein population size, which is described in detail in Section 2.2.). In an effort to obtain the most accurate results possible, only proteins with completely defined primary sequences were considered eligible for our analysis. Proteins with undetermined/unknown, ambiguous, and/or unique amino acids (B, J, O, U, X, Z) were excluded on the basis that the handling of these residues varies greatly among disorder prediction algorithms. A summary of the eligible and ineligible protein populations is displayed in Table 1. For a complete list of UniProt accession numbers for all eligible and ineligible proteins, please refer to Supplemental Table 1.

### 2.2. Sequence redundancy and uncertainty reduction

While the aforementioned eligibility screening procedure filtered out sequences that are incompatible for disorder prediction, redundant sequences and sequences that are uncertain to exist still remained in the eligible sequence population for each proteome. In order to minimize redundancy and uncertainty within the population of eligible sequences we conducted the following two-step procedure. First, UniRef100 reference cluster information was obtained via the UniProt identification mapping service (accessed programmatically on January 6, 2016) and was used to remove redundant sequences from each proteome [20, 21]. The resulting proteome populations were comprised of (i) UniProt accession numbers of eligible proteins that correspond to the unique set of UniRef100 records found to map *directly* to the reference proteome file, and (ii) UniProt accession numbers of eligible proteins that were found to map to a UniRef100 record that was not contained within the specific reference proteome file. Second, proteins with a UniProt protein existence qualifier of five were subsequently removed, as the existence of these proteins is uncertain [20]. The final population sizes have been displayed for each proteome in Table 2 (the population size of each proteome was used to determine presentation order, with *Zea mays* representing the largest population in our study following the reduction procedure). The UniProt accession numbers comprising the final population have been included in

**Table 1**
Eligibility screening summary of proteins in each studied proteome.

| Organism | Initial Total | Eligible | Ineligible |
|---|---|---|---|
| *S. cerevisiae* | 6,721 | 6,721 (100%) | 0 (0%) |
| *D. discoideum* | 12,746 | 12,733 (99.82%) | 13 (0.18%) |
| *C. reinhardtii* | 14,337 | 14,319 (99.87%) | 18 (0.13%) |
| *D. melanogaster* | 22,024 | 21,673 (98.40%) | 351 (1.60%) |
| *C. elegans* | 26,163 | 26,161 (99.99%) | 2 (0.01%) |
| *A. thaliana* | 31,551 | 31,548 (99.99%) | 3 (0.01%) |
| *D. rerio* | 41,001 | 38,192 (93.15%) | 2,809 (6.85%) |
| *M. musculus* | 45,263 | 42,306 (93.47%) | 2,957 (6.53%) |
| *H. sapiens* | 68,485 | 61,423 (89.69%) | 7,062 (10.31%) |
| *Z. mays* | 58,493 | 58,455 (99.94%) | 38 (0.06%) |

Primary sequences were obtained from UniProt reference proteome files. Proteins with undetermined, ambiguous, and/or rare amino acid residues were excluded from our analysis. Initial total, included, and excluded protein sequence counts are displayed for each organism, as well as the percentages of the initial total that have been included and excluded.