Quantitative proteome-based standards for intrinsic disorder characterization

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

[a] Department of Molecular & Integrative Physiology[1], University of Michigan Medical School, Ann Arbor, MI, USA
[b] Department of Computational Medicine & Bioinformatics[2], University of Michigan Medical School, MI, USA
[c] Department of Computational Medicine & Bioinformatics[3], 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

Running title: Standards for disorder characterization

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 standards 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 standards presented here are intended to improve the interpretation of intrinsic disorder protein content and continuous disorder predictions.

1. Introduction

Once translated, many nascent unfolded polypeptides fold into a highly ordered conformation. However, within the last two decades it 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 (IDPs) 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 IDPs as lacking function would be easy due to the absence of a well-defined tertiary structure, a growing body of evidence supports IDPs 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 1.5X 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 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 standards, specific to both proteomes and disorder

prediction tools, for gauging the significance of various disorder features. Without these standards in hand, it remains very difficult to understand whether or not a given disorder measure is significant. 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 statistical analysis must be conducted.

Motivated by these considerations, we analyzed disorder in the proteomes of ten eukaryotic model organisms using a non-parametric statistical approach. Disorder was estimated using two reputable disorder prediction algorithms, IUPred and DisEMBL, which have a physicochemical basis (single predictors will be referred to in general as "component" predictors). 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 standards 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 [17]. 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, Zea mays,* and *Homo sapiens* proteomes were included in our investigation (proteome presentation order was decided by protein population size; **Table 1**). In an effort to obtain the most accurate results possible, only proteins with completely defined primary sequences were included in 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 protein populations analyzed is displayed in

Table 1. For a complete list of all included and excluded proteins, please refer to **Supplemental Table 1**.

Table 1
Summary of the protein populations examined.

Organism	Initial Total	Included	Excluded
Saccharomyces cerevisiae	6,721	6,721 (100%)	0 (0%)
Dictyostelium discoideum	12,746	12,733 (99.82%)	13 (0.18%)
Chlamydomonas reinhardtii	14,337	14,319 (99.87%)	18 (0.13%)
Drosophila melanogaster	22,024	21,673 (98.40%)	351 (1.60%)
Caenorhabditis elegans	26,163	26,161 (99.99%)	2 (0.01%)
Arabidopsis thaliana	31,551	31,548 (99.99%)	3 (0.01%)
Danio rerio	41,001	38,192 (93.15%)	2,809 (6.85%)
Mus musculus	45,263	42,306 (93.47%)	2,957 (6.53%)
Zea mays	58,493	58,455 (99.94%)	38 (0.06%)
Homo sapiens	68,485	61,423 (89.69%)	7,062 (10.31%)

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.

2.2. Disorder prediction and analysis

Residue-specific disorder scores were obtained using the IUPred long [18, 19] and DisEMBL [20] *ab initio* disorder prediction algorithms. Due to the DisEMBL prediction of COILS (DisEMBL-C) being an overestimate of disorder (as described by [20]), only results from the HOTLOOPS (DisEMBL-H) and REM465 (DisEMBL-R) were analyzed for DisEMBL (however, DisEMBL-C probability densities have been included in **Supplemental Fig. 1, 3,** and **4**). Each residue was classified as either "ordered" or "disordered" using algorithm-specific threshold values [18-20]. Disorder was characterized in each proteome by assessing the disorder content and continuous disorder (CD) distributions. Percent disorder was calculated as the percentage of disordered residues in a protein divided by the protein length, multiplied by one hundred. A CD segment was defined as any stretch of two or more consecutive amino acids having disorder scores above the algorithm-specific threshold value.

2.3. Statistical methods

Due to the lack of normality in many of the distributions examined (**Supplemental Fig. 1, 3,** and **4**), expected values were obtained non-parametrically. Kernel density estimation (KDE) with renormalization was used to approximate the probability density function (PDF); the PDF approximation is based on the method of Jones [21]. For distributions of percentages, the PDF was approximated on the bounded domain of (0, 100). For non-percentage data, a bounded domain defined by the minimum and maximum values of the data set was used to approximate the PDF.

To determine the expected value (*E(x)*), the PDF (*f(x)*) was integrated via Eq.1 ('lb' and 'ub' represent the lower and upper bound, respectively):

$$E(x) = \int_{lb}^{ub} x f(x) dx \qquad \text{(Eq. 1)}$$

The interquartile range (25^{th} and 75^{th} percentiles) was used to examine the dispersion of the data. Additionally, in general we interpret disorder content below the 25^{th} percentile to be significantly ordered, with features above the 75^{th} percentile to be significantly disordered. However, this approach cannot be used to interpret order in IUPred, due to the large dispersion observed in its predicted disorder content distributions, which confound interpretations in the first and second quartiles. In order to provide cutoffs for gauging extreme order and disorder, the 5^{th} and 95^{th} percentile values have been reported as well.

2.4. Computational analysis

All internal noncommercial software created for use in our investigation was written in Python 2.7.10. Results were stored in SQLite databases. Data is available upon request.

3. Results

3.1. Disorder content varies approximately from 14-40% depending on predictor, with the majority of proteomes having expected values below 30%.

In order to obtain an overall summary of disorder, we first assessed the distribution of disorder percentages in each proteome. To accomplish this, disorder scores were obtained for all

residues using each of the aforementioned disorder prediction algorithms. Percent disorder was calculated as described in the **Materials and Methods**. Kernel density estimation was used to approximate the probability density function, which was then integrated using **Eq. 1** to obtain the expected value.

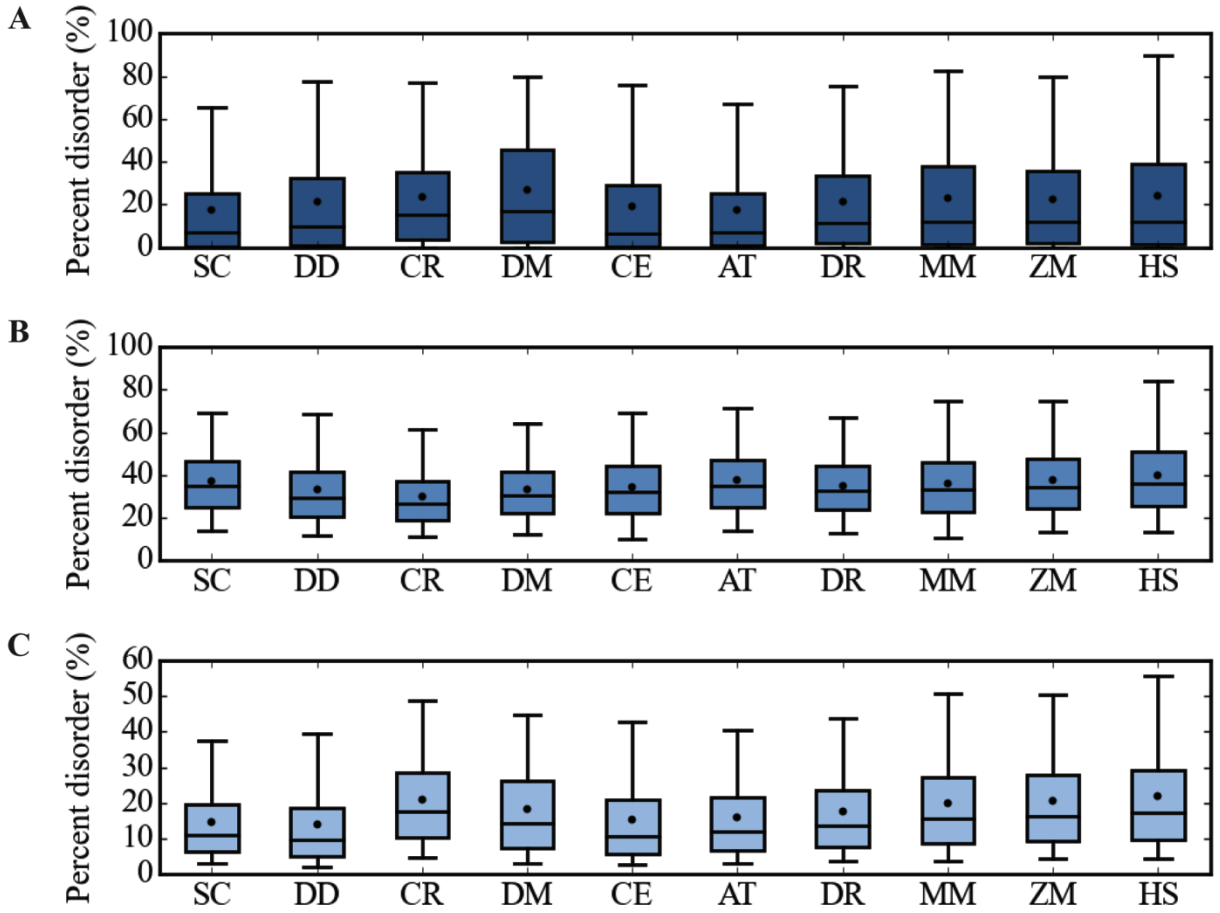

Fig. 1. Percent disorder distribution in ten eukaryotic proteomes. Boxplots of percent disorder determined by (A) IUPred, (B) DisEMBL-H, and (C) DisEMBL-R are shown. The horizontal line indicates the median, whereas the dots indicate the expected value determined via **Eq. 1**. The whiskers represent the 5th and 95th percentile values. Numerical values are summarized in **Table 3**.

The disorder content distribution is positively skewed in each of the ten eukaryotes analyzed. Expected values were found to range between ~17-27%, ~30-40%, and ~14-22% for IUPred, DisEMBL-H, and DisEMBL-R, respectively (**Fig. 1**). Whereas the human disorder content distribution exhibited the greatest dispersion for DisEMBL predictions (**Fig. 1B, C**), the

IUPred distributions did not follow this trend, as the greatest spread was found in the *Drosophila melanogaster* proteome (**Fig. 1A**). Nevertheless, the IUPred and DisEMBL-R component predictions were consistent with the 20.5% average disorder percentage recently reported by an investigation of disorder in 110 eukaryotes using IUPred and Espritz [16], as well as an earlier proteomic study conducted using DISOPRED2 that found disorder content to vary from ~16-22% in five eukaryotes [14]. DisEMBL-H predictions were found to be much higher overall, but still in agreement with the 35-45% range reported by [15], which utilized the PONDR VSL2B predictor. Probability densities are displayed in **Supplemental Fig. 1**.

3.2. The majority of the studied eukaryotes contain a least one continuous disorder domain.

By definition, a CD region must contain a minimum of two consecutive disordered amino acids. With a length of two amino acids representing the lower bound of integration when applying **Eq. 1** to determine the expected length of a CD segment, completely ordered proteins (0% disorder) and proteins with disorder composition consisting entirely of isolated disordered amino acids must be excluded from our CD analysis. Over 70% of all eukaryotic proteomes contain at least one CD stretch as determined by any of the single component predictors (**Table 2**). Moreover, disorder percentages within the CD-containing protein populations were nearly identical to those of the entire populations for DisEMBL-H and DisEMBL-R component predictions (compare **Supplemental Fig. 2A** to **Fig. 1**), whereas minors differences were observed for IUPred-predicted disorder (**Supplemental Fig. 2A, B**). This result can be explained by the fact that IUPred predicted a greater amount of isolated disordered amino acids than did DisEMBL, causing the size of the total and CD-containing protein populations to be substantially different for IUPred and identical for DisEMBL (**Table 2**). Nevertheless, provided over two thirds of each proteome exhibited continuous disorder, we reasoned that all of the eukaryotes included in our investigation contained a population of eligible (CD-containing) proteins large enough to determine representative expected values for various CD features.

3.3. For many of the eukaryotes examined, a CD region greater than 30 amino acids is expected, and the length of significantly long disordered stretches varies substantially between predictors.

Isolated disordered amino acids and continuous clusters of disordered residues constitute the two most basic disorder arrangements. While isolated disordered residues may influence the

194 structure of some proteins, longer continuously disordered segments provide better indicators of
195 protein regions that are more strongly influenced by disorder. In previous studies, CD prevalence
196 has often been assessed by estimating the percentage of a proteome containing a CD stretch
197 greater than or equal to 30 amino acids in length [9, 14-16, 22, 23]. However, to our knowledge
198 rigorously determined expected values and detailed ranges have not been reported numerically.
199 Here, we assessed the distribution of the longest CD region (CD_L) in each proteome.
200

Table 2
Summary of continuous disorder prevalence in ten eukaryotic proteomes.

Organism	IUPred	DisEMBL-H	DisEMBL-R
Saccharomyces cerevisiae	4,994 (74.30%)	6,718 (99.96%)	6,710 (99.84%)
Dictyostelium discoideum	9,857 (77.41%)	12,728 (99.96%)	12,710 (99.82%)
Chlamydomonas reinhardtii	12,184 (85.09%)	14,317 (99.99%)	14,314 (99.97%)
Drosophila melanogaster	18,217 (84.05%)	21,667 (99.97%)	21,670 (99.99%)
Caenorhabditis elegans	18,536 (70.85%)	26,152 (99.97%)	26,122 (99.85%)
Arabidopsis thaliana	24,571 (77.88%)	31,544 (99.99%)	31,539 (99.97%)
Danio rerio	31,422 (82.27%)	38,187 (99.99%)	38,178 (99.96%)
Mus musculus	33,035 (78.09%)	42,294 (99.97%)	42,299 (99.98%)
Zea mays	47,109 (80.59%)	58,422 (99.94%)	58,427 (99.95%)
Homo sapiens	46,505 (75.71%)	61,411 (99.98%)	61,400 (99.96%)

CD regions were defined as any stretch of consecutively disordered amino acids greater or equal than 2 amino acids in length. The number of proteins containing the CD feature is displayed along with the percentage of the included proteome population in parentheses (see **Table 1** for the included population descriptions). Results are shown for all component disorder prediction algorithms.

201

202 For IUPred, DisEMBL-H, and DisEMBL-R component predictors, the CD_L expected
203 values varied from ~20 amino acids (*S. cerevisiae* and *A. thaliana*, DisEMBL-R) to ~103 amino
204 acids (*D. melanogaster*, IUPred), with 19 of 30 expected values being greater than or equal to 30
205 amino acids (**Fig. 2**). Segment lengths of seven amino acids (*D. discoideum*, DisEMBL-R) and
206 132 amino acids (*D. melanogaster*, IUPred) represented the respective lowest 25th percentile
207 value and the greatest 75th percentile value for CD_L stretches estimated via component
208 predictions (**Fig. 2**). Interestingly, the dispersion of the IUPred distributions was far greater than
209 those of DisEMBL-H and DisEMBL-R. 39 amino acids (*A. thaliana*) and 117 amino acids (*D.*
210 *melanogaster*) were found to be the minimum and maximum IQR size for CD_L distributions
211 predicted by IUPred (**Fig. 2A**), whereas the respective minimum-maximum IQR sizes for

DisEMBL-H and DisEMBL-R were 15 (*C. reinhardtii*)-25 amino acids (*D. melanogaster*) and 16 amino acids (*A. thaliana*)-30 amino acids (*D. melanogaster*) (**Fig. 2B, C**). While the greatest 95th percentile values for CD_L varied greatly from 50 amino acids (DisEMBL-R, *A. thaliana*) to 353 amino acids (IUPred, *D. melanogaster*), all (30 out of 30) were above 50 amino acids. Probability densities are shown in **Supplemental Fig. 3**.

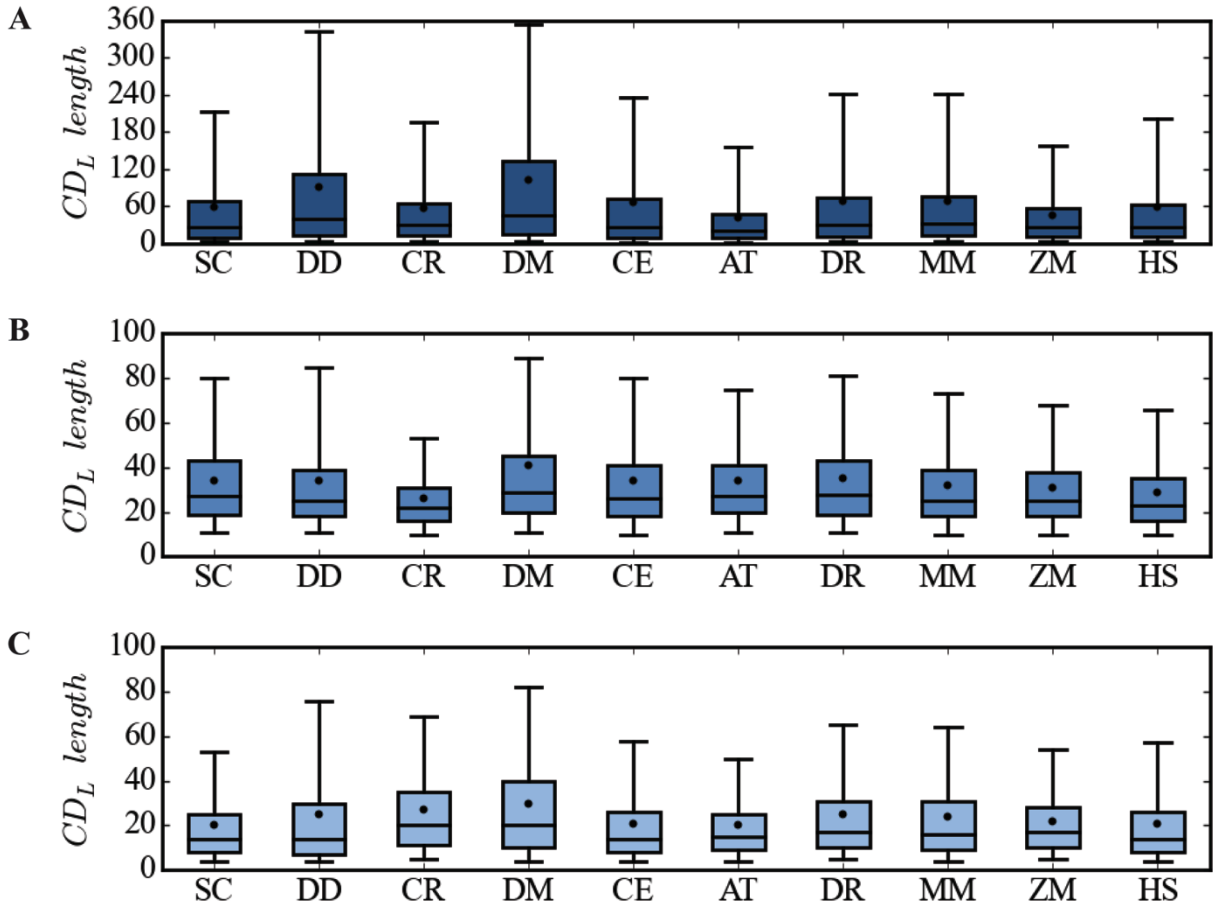

Fig 2. Longest CD stretch distribution in ten eukaryotic proteomes. Boxplots of CD_L regions determined by IUPred (A), DisEMBL-H (B), and DisEMBL-R (C) are shown. The horizontal line within the box indicates the median, whereas the dots indicate the expected value determined via **Eq. 1**. The whiskers indicate the 5th and 95th percentile values. Numerical values are summarized in **Table 3**.

For the component predictors considered here, our results indicate that the significance threshold of 30 amino acids may only be appropriate when using DisEMBL-R, as this length often fails exceed the interquartile range for DisEMBL-H and IUPred predictions. Provided eight of the ten eukaryotes had a 75th percentile for the CD_L region greater than or equal to 38 amino

acids and 60 amino acids for DisEMBL-H and IUPred predictors, respectively, predictor-specific threshold values greater than the currently used 30 aa CD_L significance value should be established for these predictors (see **Discussion**). Furthermore, the differences in the magnitude of both the expected and 75th percentile values observed here underscore the need to establish predictor-specific thresholds instead of adhering to a single, universal value.

3.4. In CD-containing proteins, the longest disordered region accounts for 6-19% of a protein's total length.

While the results presented in **Fig. 2** provide useful, intuitive information regarding the typical length expected of the CD_L region contained within a protein, it is subject to nebulous interpretations as it is not in the context of primary sequence length. To address this issue, we next characterized the percentage of the total length of a protein that is accounted for by the longest continuously disordered segment. For each CD-containing protein, the longest CD percentage of length (LCPL) was simply calculated by dividing the length of the CD_L by the primary sequence length and multiplying the result by one hundred. LCPL distributions were analyzed in each proteome and expected values were obtained via **Eq. 1**.

The expected values for LCPL varied from ~6% (*S. cerevisiae*, DisEMBL-R) to 19% (*H. sapiens*, IUPred), with minimum 25th and maximum 75th percentile values of ~2% (*D. discoideum* and *C. elegans*, DisEMBL-R) and 25% (*H. sapiens*, IUPred), respectively (**Fig. 3**). The organisms defining the boundaries of the LCPL ranges also exhibited the least (*S. cerevisiae*, DisEMBL-R) and most (*H. sapiens*, IUPred) dispersion within the central 50% of the population (**Fig. 3**). Furthermore, we found the 95th percentile for LCPL to vary dramatically from roughly 18% (*S. cerevisiae* and *D. melanogaster*, DisEMBL-R) to ~82% (*H. sapiens*, IUPred) (**Fig. 3**). Provided all of the expected values, and 24 out of 30 of the 75th percentile values are below 20%, our results suggest the CD_L segment contained in a protein is typically less than 20% of the total protein length, and CD segments occupying a greater percentage of the total length may be significant. Probability densities can be found in **Supplemental Fig. 4**.

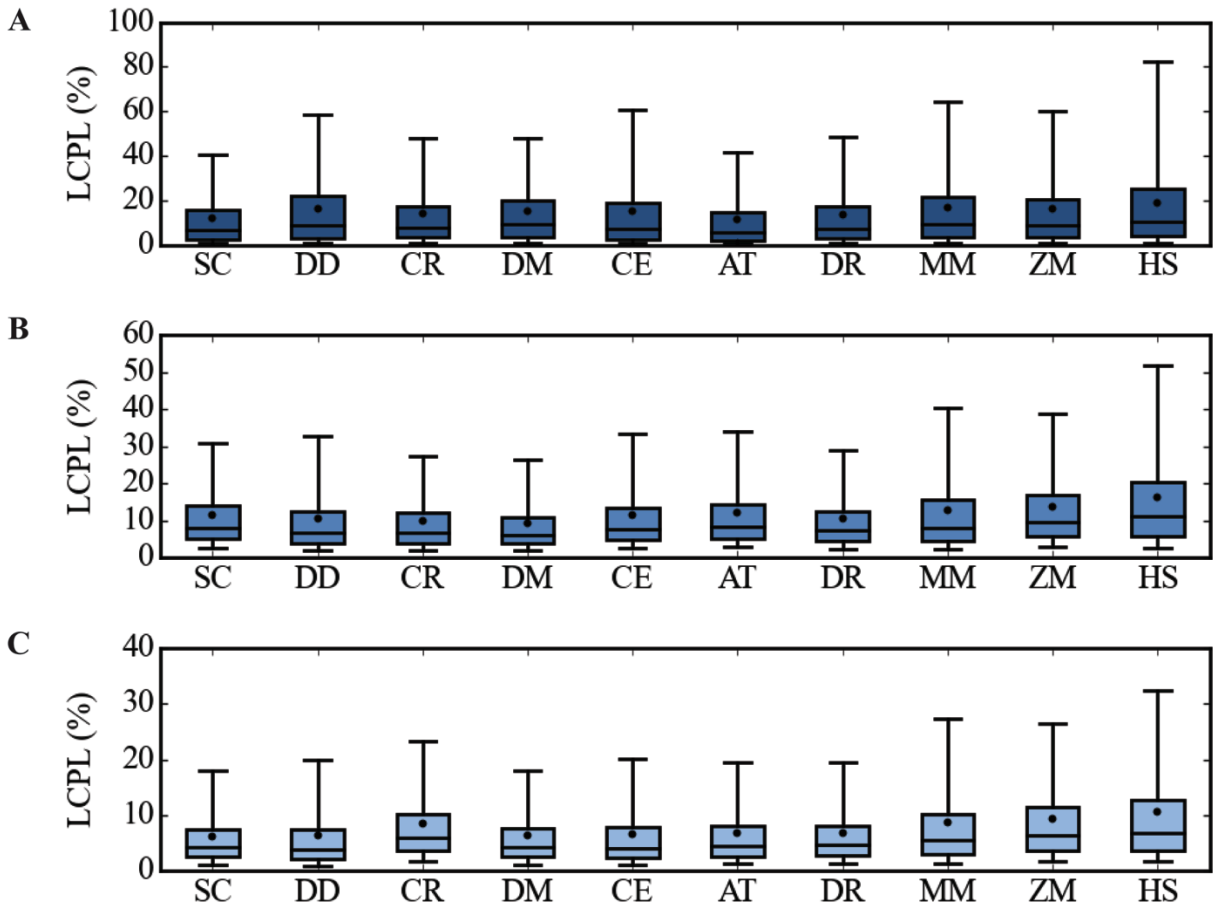

Fig 3. Longest CD percentage of length (LCPL) distribution in ten eukaryotic proteomes. Boxplots of LCPL determined by IUPred (A), DisEMBL-H (B), and DisEMBL-R (C) are shown. The horizontal line within the box indicates the median, whereas the dots indicate the expected value determined via **Eq. 1**. The whiskers represent the 5th and 95th percentiles. Numerical values are summarized in **Table 3**.

3.5. The reliability of significance thresholds for CD length varies between predictors in a protein length-dependent fashion.

When assessing the prevalence of significantly long CD regions, the percentage of CD segments greater than or equal to a fixed length is often examined, with 30 amino acids representing the commonly used value [9, 14-16, 22, 23]. However, when deeming a CD region as significant on the basis of length alone, utilizing a fixed threshold length becomes less reliable as the primary sequence length increases. For instance, many would be more willing to accept a 30 amino acids long CD region as being significant in a protein that is 300 amino acids long compared to the same length segment in a 3,000 amino acids long protein, as this region

accounts for a greater percentage of the length in the former protein and is arguably more likely to hold greater influence over structure and function overall. This consideration leads us to arrive at the following question. When is a protein too long to evaluate the significance of a CD region on the basis of its length alone?

To answer this question, we determined the protein length at which the CD_L region expected value (**Fig. 2**) begins to fall below the 25th percentile cutoff for the LCPL **(Fig. 3**) (the concept of this threshold protein length is depicted in **Fig. 4A**). These values were obtained by solving **Eq. 2** for the 25th percentile LCPL value:

$$LCPL^* = \frac{CD_L\, E(X)}{PL} * 100 \qquad \text{(Eq. 2)}$$

In **Eq. 2**, CD_L E(X) is the expected value determined for the longest CD region and PL is the primary sequence length. **Fig. 4A** illustrates the concept used for a single predictor in a single proteome, whereas the results for all predictors in the CD-containing proteins of all proteomes are displayed in **Fig. 4B**.

Protein length threshold (PLT) values ranged from 1,314-2,943 amino acids, 483-1,051 amino acids, and 568-1,200 amino acids for IUPred, DisEMBL-H, and DisEMBL-R, respectively (**Fig. 4B**). In all cases, PLT values were found to be substantially longer for IUPred compared to DisEMBL (**Fig. 4B**). For DisEMBL predictions, the longest DisEMBL-R-determined CD expected value was more tolerant of longer proteins than the corresponding DisEMBL-H value (**Fig. 4B**). While the percentage of proteomes (specifically, the CD-containing population of each proteome) consisting of proteins with a primary sequence length greater than the predictor-specific PLT was low for IUPred (due to higher E(X)), it was substantially greater for DisEMBL predictions, as it exceeded 10-15% in many cases (**Fig. 4C**). Thus, this result suggests that while the deficiency inherent to the CD_L metric may be of little concern when using IUPred, greater attention must be given to protein length when assessing CD significance using DisEMBL.

3.6. LCPL is stricter than CD_L when gauging the significance of continuous disordered regions in long proteins.

We subsequently compared the effect of two different length thresholds (LCPL and PLT), the commonly used 30 amino acids value and the CD_L 75th percentile values (**Fig. 2**), as well as the LCPL 75th percentile values (**Fig. 3**), in identifying proteins with a CD region of

significant length. Specifically, this analysis was performed in the subpopulation of CD containing proteins with a primary sequence length greater than or equal to the prediction algorithm-specific threshold primary sequence length determined for each proteome (**Fig. 4B**).

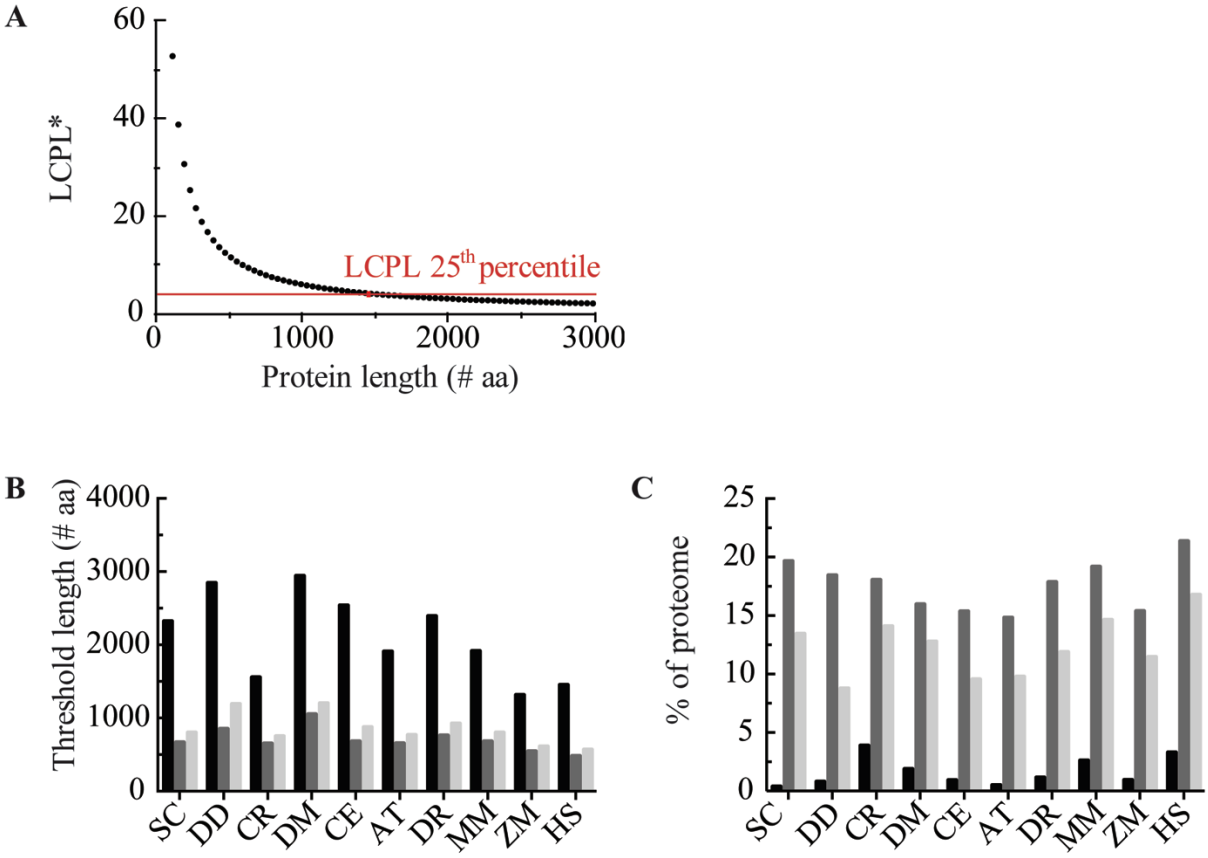

Fig. 4. Reliability of CD length thresholds with increasing protein length. (A) Graphical depiction of the concept for estimating LCPL protein length thresholds. **Eq. 2** was solved to find the protein length where the LCPL begins to fall below the 25th percentile LCPL value specific to the proteome and prediction algorithm. The red dot is the IUPred protein length threshold result in the *Homo sapiens* proteome. The black dots are LCPL values calculated with Eq. 2 using protein lengths from 110 to 3,000 amino acids and are for conceptual purposes only. (B) Protein length threshold (PLT) values marking the maximum protein length where CD regions can be considered significant on the basis of length alone. Numeric values are provided in **Table 3**. (C) Percentage of the CD-containing proteins of each proteome with a length greater than or equal to the threshold values displayed in (B). Results for all proteomes are displayed for IUPred (black), DisEMBL-H (light gray), DisEMBL-R (dark gray).

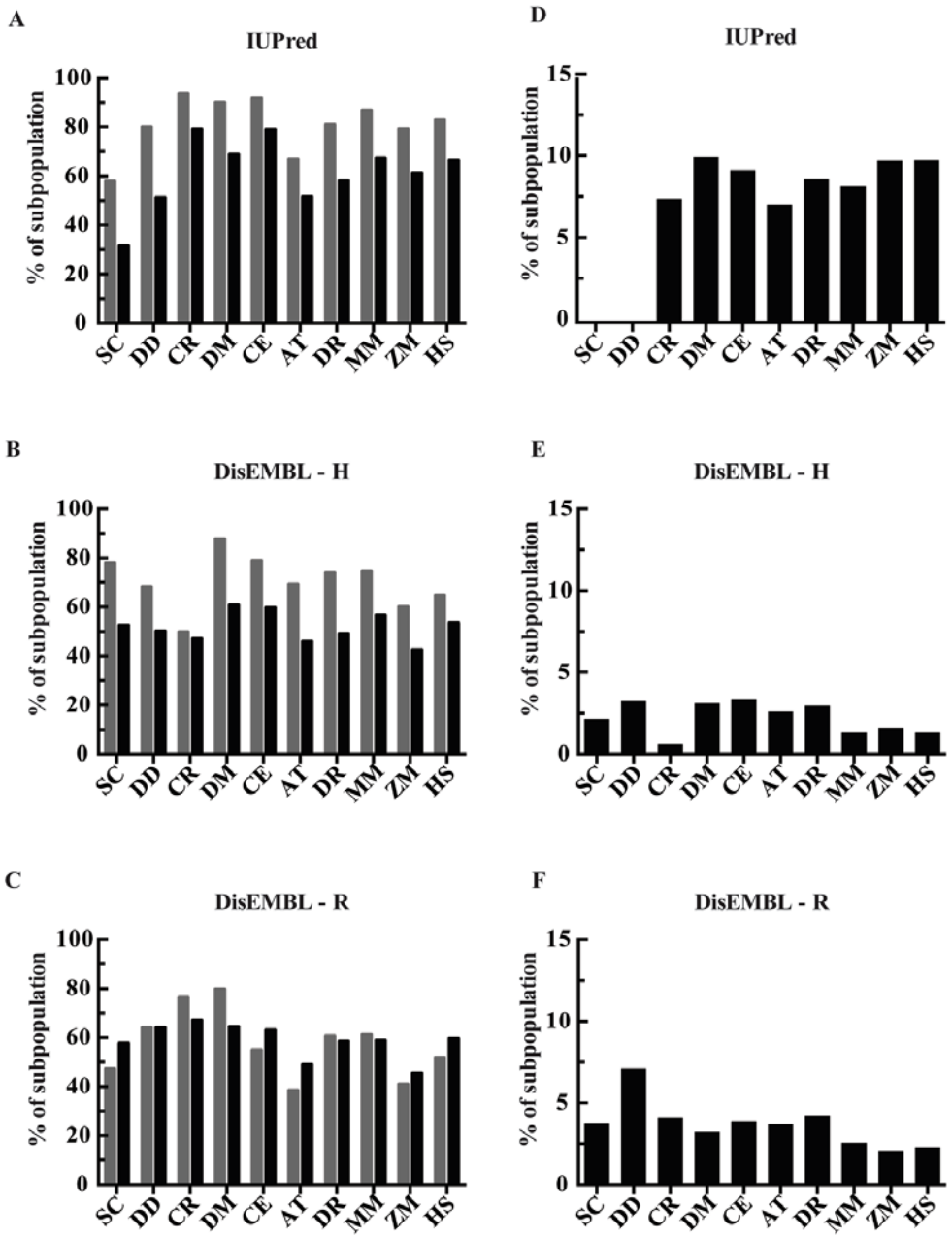

Fig. 5. LCPL is a strict metric for gauging the significance of continuously disordered regions in long proteins. The subpopulation of proteins having a length greater than or equal to the predictor-specific protein length threshold presented in **Fig. 4** was analyzed. (A-C) The percentage of the subpopulation containing a longest continuous disordered region greater than or equal to 30 amino acids (gray bars), or the proteome-specific 75th percentile value (black bars). (D-F) The percentage of the subpopulation having a LCPL value greater than or equal to the proteome-specific 75th percentile value. Results are displayed for IUPred (A, D), DisEMBL-H (B, E), and DisEMBL-R (C, F).

Between the two length thresholds, a smaller percentage of each subpopulation was predicted to contain a significantly long disordered segment when using the LCPL 75th percentile determined in this study (**Fig. 5A, B**), although this was more variable with DisEMBL-R due to the more conservative nature of its CD predictions (**Fig. 5C**). Nevertheless, a substantial fraction of each subpopulation was still found to contain the feature of interest (**Fig. 5A-C**), supporting the **Fig. 4** assertion that conferring significance to CD region on the basis of raw length alone is inappropriate for proteins with a residue count exceeding the algorithm-specific PLT. We subsequently explored the effect of classification using the LCPL metric. From **Fig. 5D-F**, a substantial decrease in the percentage of proteins exhibiting a significant CD_L segment is observed when using the LCPL for classification. In all cases, less than 15% of the length threshold-exceeding subpopulation was classified as having a significantly long CD region, whereas the same was found for less than 10% of the DisEMBL subpopulations (**Fig. 5D-F**). Taken together, these results exemplify the point that the LCPL is the superior metric for evaluating significantly long disordered regions in proteins having lengths exceeding the predictor-specific threshold values determined in **Fig. 4**.

4. Discussion

We presented a thorough analysis of intrinsic disorder predicted by two reputable algorithms based on physicochemical principles. Our analysis utilized a non-parametric statistical approach to estimate objective standards for determining whether the disorder content or length of a CD region is significant. A summary of our standards is displayed in **Table 3**.

Expected values and ranges were found to vary between the different component predictors. DisEMBL-R, an artificial neural network trained on missing electron density assignments in the Protein Data Bank, consistently predicted fewer disordered residues than did IUPred or DisEMBL-H (**Fig. 1**). Overall, the disorder content ranges for IUPred and DisEMBL-R were in agreement with a range of ~16-22% [14] and an average of 20.5% [16] reported by two proteomic investigations of intrinsic disorder, whereas DisEMBL-H disorder content predictions were found to be substantially higher, but still consistent with results from a third proteomic disorder investigation [15] (**Fig. 1**). Moreover, although DisEMBL-H predicted more disorder than DisEMBL-R, it does so with greater accuracy given its lower false positive rate [20], as disorder only represents one potential cause for missing coordinates in the X-ray

crystallography data used to train DisEMBL-R. Regarding the IUPred predictions, it is important to understand that the great dispersion observed in the IUPred disorder content distributions (**Fig. 1A**) prevents the use of its 25th percentile as a mark above which proteins begin to lack significant order. Regardless, the 25th percentile values are interpretable for DisEMBL predictions, and in all cases, the 75th percentile values can be used to define proteins containing significantly high disorder content.

To provide insight into the organization of disordered residues, the distributions of the longest CD region (CD_L) and the percentage of residues accounted for by the longest continuous disordered region (LCPL) were examined. Due to the restraints of **Eq. 1**, we limited all CD analyses to the CD-containing population of each proteome (**Table 2**), resulting in the exclusion of completely ordered proteins, as well as proteins exhibiting disorder exclusively in the form of isolated amino acids. While this may lead to inflation of the disorder features with respect to the whole proteome population, a large fraction of each proteome (>70% in all cases) was found to exhibit CD, as predicted by each algorithm (**Table 2**), therefore leading us to accept these populations as being representative and dismiss any inflation resulting from the enforcement of the eligibility criterion to be minimal. Additionally, we acknowledge that the CD_L metric does not account for various features contributing to the overall disorder content of a protein, such as isolated disordered amino acids and/or shorter CD segments that may exist in areas nearby the CD_L region or other CD regions. Regardless, CD_L provides an indication of the most organized disordered segment in a protein, and when examined at the population level, it offers valuable insight into the largest disordered segment one should expect in a given protein. IUPred predictions provided the greatest expected values with the most dispersion (**Fig. 2A**) when examining CD_L regions; this trend was also found in the LCPL distributions (**Fig. 3A**).

When deeming CD regions significant, it has become common practice to make this evaluation with respect to a fixed length. Numerous studies utilizing various component predictors have most commonly used a threshold value of 30 consecutive disordered residues to define a 'long disordered region' [9, 14-16, 22, 23]. For DisEMBL-R, this value appears to be a reasonable cutoff for significance, as none of the proteomes were found to have 75th percentile cutoffs nor expected values greater than 30 amino acids (**Table 3**). However, the threshold of 30 amino acids appears to be insufficient for identifying significantly long CD_L regions when using the IUPred and DisEMBL-H algorithms. This was suggested by the observation that all

proteomes were found to have CD_L 75th percentile values greater than 30 amino acids when assessing CD_L regions with IUPred and DisEMBL - H. Furthermore, all and eight of ten proteomes were found to contain expected values for CD_L length greater than 30 amino acids for IUPred and DisEMBL-H, respectively.

Table 3
Summary of disorder standards.

		PD	CD_L	LCPL	PLT
S. cerevisiae	**I**	17.4 (0.6-25.1)	58 (9-68)	11.9 (2.5-15.5)	2320
	H	37.1 (25.0-46.3)	34 (19-43)	11.6 (5.1-14.0)	667
	R	14.5 (6.3-19.4)	20 (8-25)	6.3 (2.5-7.4)	800
D. discoideum	**I**	21.1 (1.1-32.2)	91 (13-112)	16.3 (3.2-21.9)	2844
	H	33.2 (20.8-41.5)	34 (18-39)	10.7 (4.0-12.7)	850
	R	14.0 (5.0-18.7)	25 (7-30)	6.4 (2.1-7.4)	1190
C. reinhardtii	**I**	23.3 (3.9-35.1)	56 (12-65)	13.9 (3.6-17.0)	1556
	H	30.1 (19.0-37.4)	26 (16-31)	9.9 (4.0-12.2)	650
	R	21.0 (10.3-28.4)	27 (11-35)	8.5 (3.6-10.3)	750
D. melanogaster	**I**	26.8 (2.8-45.4)	103 (15-132)	15.0 (3.5-20.0)	2943
	H	33.4 (22.3-41.9)	41 (20-45)	9.3 (3.9-10.8)	1051
	R	18.3 (7.3-26.2)	30 (10-40)	6.4 (2.5-7.6)	1200
C. elegans	**I**	19.1 (0.1-29.0)	66 (9-72)	15.1 (2.6-18.8)	2538
	H	34.7 (22.1-44.5)	34 (18-41)	11.5 (5.0-13.5)	680
	R	15.4 (5.8-20.8)	21 (8-26)	6.7 (2.4-7.8)	875
A. thaliana	**I**	17.2 (1.0-25.2)	42 (8-47)	11.6 (2.2-14.5)	1909
	H	37.6 (25.2-47.1)	34 (20-41)	12.1 (5.2-14.6)	654
	R	15.9 (6.8-21.5)	20 (9-25)	6.8 (2.6-8.2)	769
D. rerio	**I**	21.5 (2.0-33.6)	67 (11-73)	13.6 (2.8-17.3)	2393
	H	35.2 (23.7-44.1)	35 (19-43)	10.6 (4.6-12.6)	761
	R	17.5 (7.6-23.6)	25 (10-31)	6.9 (2.7-8.1)	926
M. musculus	**I**	23.3 (1.5-37.9)	67 (12-76)	16.5 (3.5-21.5)	1914
	H	36.1 (23.1-46.0)	32 (18-39)	13.0 (4.7-15.6)	681
	R	20.0 (8.5-27.2)	24 (9-31)	8.8 (3.0-10.2)	800
Z. mays	**I**	22.5 (2.1-35.4)	46 (11-57)	16.0 (3.5-20.6)	1314
	H	37.8 (24.6-47.9)	31 (18-38)	13.7 (5.7-16.9)	544
	R	20.5 (9.2-27.8)	22 (10-28)	9.3 (3.6-11.5)	611
H. sapiens	**I**	23.8 (1.2-38.9)	58 (10-63)	19.0 (4.0-25.0)	1450
	H	39.8 (25.4-51.2)	29 (16-35)	16.5 (6.0-20.5)	483
	R	21.9 (9.8-29.3)	21 (8-26)	10.6 (3.7-12.8)	568

The expected values for percent disorder (PD), the longest continuous disordered region (CD_L), CD_L percentage of length (LCPL), and the protein length threshold (PLT) beyond which a CD region cannot be determined significant on the basis of length alone are presented for the three disorder prediction algorithms and all ten proteomes included in this investigation. The interquartile range is presented in parentheses. I, H, and R represent IUPred, DisEMBL – H, and DisEMBL – R, respectively.

Considering the above, we suggested algorithm-specific thresholds be established that extend beyond the current 30 amino acids value, when conferring significance to CD_L regions on the basis of length alone. For DisEMBL-H, we propose that a CD_L threshold length of 40 amino acids would be more appropriate on the basis that eight of the ten eukaryotes exhibited 75^{th} percentile values greater or equal than 38 amino acids, with five of ten being exceeding 40 amino acids. For IUPred, the threshold value should be set even higher, at 60 amino acids, as eight of the ten proteomes had 75^{th} percentile values greater than 60 amino acids (albeit the range of the IUPred CD_L 75^{th} percentile values were much greater than those of DisEMBL-H) (**Table 3**). The proposed increases would be substantial, with the DisEMBL-H and IUPred thresholds representing a 33% and 100% increase over the existing value of 30 amino acids.

One major concern of the CD_L metric is that its predictive power diminishes with increasing primary sequence length. To provide a means for identifying when a protein is too long for compatibility with this metric, we estimated threshold protein lengths (referred to as a "PLT"), specific to each disorder prediction algorithm, and for all ten eukaryotes analyzed (**Fig. 4**). The prevalence of this issue was dramatically lower for IUPred predictions, given less than 5% of every proteome had a length greater than the predictor-specific PLT; whereas the issue was much more pronounced for DisEMBL predictions as over 10% of most proteomes were found to exceed the PLT length (**Fig. 4C**). When assessing the significance of a CD region in a protein exceeding these aforementioned threshold protein lengths, it is recommended that the LCPL metric be used in place of a length threshold, as we have shown the LCPL to be far more selective in general (**Fig. 5**).

In closing, the standards presented here are intended to facilitate biochemical and biophysical scientists in making objective disorder classifications in a protein of interest belonging to one of the ten eukaryotic proteomes included in our analysis. Although our study was limited to a smaller number of prediction tools, the general analytical approach is amenable to any disorder prediction algorithm with computational performance suitable for whole proteome analysis. Thus, a bigger picture goal of this work is that it will inspire similar analyses to be performed prior to the release of new disorder prediction algorithms, as well as for other existing algorithms, in order to facilitate the interpretation of disorder predictions. With a universal disorder prediction tool currently absent, together with the variation in disorder

423 predictions observed between different algorithms and between different proteomes, the
424 meaningful interpretation of disorder relies heavily on standards like the ones presented in this
425 work.
426
427 **Acknowledgements**
428 We thank Dr. Kerby Shedden (Center for Statistical Consultation and Research, University of
429 Michigan) for his technical support. We also thank Suzanne Shoffner for critically reading the
430 manuscript. This work was partially supported by the University of Michigan Protein Folding
431 Diseases Initiative and the University of Michigan Medical School Research Discovery Fund.
432
433 **References**
434
435 [1] R.W. Kriwacki, L. Hengst, L. Tennant, S.I. Reed, P.E. Wright, Structural studies of
436 p21Waf1/Cip1/Sdi1 in the free and Cdk2-bound state: conformational disorder mediates binding
437 diversity, Proc Natl Acad Sci U S A, 93 (1996) 11504-11509.
438 [2] K.W. Plaxco, M. Grob, The importance of being unfolded, Nature, 386 (1997) 657-659.
439 [3] P.E. Wright, H.J. Dyson, Intrinsically unstructured proteins: re-assessing the protein
440 structure-function paradigm, J Mol Biol, 293 (1999) 321-331.
441 [4] H.J. Dyson, P.E. Wright, Intrinsically unstructured proteins and their functions, Nat Rev Mol
442 Cell Biol, 6 (2005) 197-208.
443 [5] P. Tompa, Intrinsically disordered proteins: a 10-year recap, Trends Biochem Sci, 37 (2012)
444 509-516.
445 [6] L.M. Iakoucheva, C.J. Brown, J.D. Lawson, Z. Obradović, A.K. Dunker, Intrinsic Disorder in
446 Cell-signaling and Cancer-associated Proteins, Journal of Molecular Biology, 323 (2002) 573-
447 584.
448 [7] P.E. Wright, H.J. Dyson, Intrinsically disordered proteins in cellular signalling and
449 regulation, Nat Rev Mol Cell Biol, 16 (2015) 18-29.
450 [8] G.J. Rautureau, C.L. Day, M.G. Hinds, Intrinsically disordered proteins in bcl-2 regulated
451 apoptosis, Int J Mol Sci, 11 (2010) 1808-1824.
452 [9] Z. Peng, B. Xue, L. Kurgan, V.N. Uversky, Resilience of death: intrinsic disorder in proteins
453 involved in the programmed cell death, Cell Death Differ, 20 (2013) 1257-1267.

454 [10] M.K. Yoon, D.M. Mitrea, L. Ou, R.W. Kriwacki, Cell cycle regulation by the intrinsically
455 disordered proteins p21 and p27, Biochem Soc Trans, 40 (2012) 981-988.
456 [11] V. Vittal, L. Shi, D.M. Wenzel, K.M. Scaglione, E.D. Duncan, V. Basrur, K.S. Elenitoba-
457 Johnson, D. Baker, H.L. Paulson, P.S. Brzovic, R.E. Klevit, Intrinsic disorder drives N-terminal
458 ubiquitination by Ube2w, Nat Chem Biol, 11 (2015) 83-89.
459 [12] B. Wang, S. Merillat, M. Vincent, A. Huber, V. Basrur, L. Zeng, K. Elenitoba-Johnson, R.
460 Miller, D. Irani, A. Dlugosz, S. Schnell, K. Scaglione, H. Paulson, Loss of the ubiquitin-
461 conjugating enzyme Ube2W results in susceptibility to early postnatal lethality and defects in
462 skin, immune and male reproductive systems., The Journal of Biological Chemistry, DOI:
463 10.1074/jbc.M115.676601 (2015).
464 [13] T.P. Knowles, M. Vendruscolo, C.M. Dobson, The amyloid state and its association with
465 protein misfolding diseases, Nat Rev Mol Cell Biol, 15 (2014) 384-396.
466 [14] J.J. Ward, J.S. Sodhi, L.J. McGuffin, B.F. Buxton, D.T. Jones, Prediction and functional
467 analysis of native disorder in proteins from the three kingdoms of life, J Mol Biol, 337 (2004)
468 635-645.
469 [15] B. Xue, A.K. Dunker, V.N. Uversky, Orderly order in protein intrinsic disorder distribution:
470 disorder in 3500 proteomes from viruses and the three domains of life, J Biomol Struct Dyn, 30
471 (2012) 137-149.
472 [16] Z. Peng, J. Yan, X. Fan, M.J. Mizianty, B. Xue, K. Wang, G. Hu, V.N. Uversky, L. Kurgan,
473 Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all
474 domains of life, Cell Mol Life Sci, 72 (2015) 137-151.
475 [17] C. UniProt, UniProt: a hub for protein information, Nucleic Acids Res, 43 (2015) D204-
476 212.
477 [18] Z. Dosztanyi, V. Csizmok, P. Tompa, I. Simon, The pairwise energy content estimated from
478 amino acid composition discriminates between folded and intrinsically unstructured proteins, J
479 Mol Biol, 347 (2005) 827-839.
480 [19] Z. Dosztanyi, V. Csizmok, P. Tompa, I. Simon, IUPred: web server for the prediction of
481 intrinsically unstructured regions of proteins based on estimated energy content, Bioinformatics,
482 21 (2005) 3433-3434.
483 [20] R. Linding, L.J. Jensen, F. Diella, P. Bork, T.J. Gibson, R.B. Russell, Protein disorder
484 prediction: implications for structural proteomics, Structure, 11 (2003) 1453-1459.

485 [21] M.C. Jones, Simple boundary correction for kernel density estimation, Statistics and
486 Computing, 3 (1993) 135-146.
487 [22] C. Haynes, C.J. Oldfield, F. Ji, N. Klitgord, M.E. Cusick, P. Radivojac, V.N. Uversky, M.
488 Vidal, L.M. Iakoucheva, Intrinsic disorder is a common feature of hub proteins from four
489 eukaryotic interactomes, PLoS Comput Biol, 2 (2006) e100.
490 [23] Z. Peng, M.J. Mizianty, L. Kurgan, Genome-scale prediction of proteins with long
491 intrinsically disordered regions, Proteins, 82 (2014) 145-158.