Biochimica et Biophysica Acta 1834 (2013) 1671–1680

Contents lists available at SciVerse ScienceDirect

Biochimica et Biophysica Acta

journal homepage: www.elsevier.com/locate/bbapap

RAPID: Fast and accurate sequence-based prediction of intrinsic disorder content on proteomic scale

Jing Yan [a], Marcin J. Mizianty [a], Paul L. Filipow [a], Vladimir N. Uversky [b,c], Lukasz Kurgan [a,*]

[a] Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Canada
[b] Department of Molecular Medicine, Byrd Alzheimer's Research Institute, Morsani College of Medicine, University of South Florida, Tampa, FL 33612, USA
[c] Institute for Biological Instrumentation, Russian Academy of Sciences, 142290 Pushchino, Moscow Region, Russia

ARTICLE INFO

Article history:
Received 25 March 2013
Accepted 22 May 2013
Available online 1 June 2013

Keywords:
Intrinsic disorder
Disorder content
Disorder prediction
Eukaryotes
Structural coverage

ABSTRACT

Recent research in the protein intrinsic disorder was stimulated by the availability of accurate computational predictors. However, most of these methods are relatively slow, especially considering proteome-scale applications, and were shown to produce relatively large errors when estimating disorder at the protein- (in contrast to residue-) level, which is defined by the fraction/content of disordered residues. To this end, we propose a novel support vector Regression-based Accurate Predictor of Intrinsic Disorder (RAPID). Key advantages of RAPID are speed (prediction of an average-size eukaryotic proteome takes < 1 h on a modern desktop computer); sophisticated design (multiple, complementary information sources that are aggregated over an input chain are combined using feature selection); and high-quality and robust predictive performance. Empirical tests on two diverse benchmark datasets reveal that RAPID's predictive performance compares favorably to a comprehensive set of state-of-the-art disorder and disorder content predictors. Drawing on high speed and good predictive quality, RAPID was used to perform large-scale characterization of disorder in 200+ fully sequenced eukaryotic proteomes. Our analysis reveals interesting relations of disorder with structural coverage and chain length, and unusual distribution of fully disordered chains. We also performed a comprehensive (using 56000+ annotated chains, which doubles the scope of previous studies) investigation of cellular functions and localizations that are enriched in the disorder in the human proteome. RAPID, which allows for batch (proteome-wide) predictions, is available as a web server at http://biomine.ece.ualberta.ca/RAPID/.

1. Introduction

Intrinsically disordered proteins and intrinsically disordered protein regions lack a unique 3-D structure, and exist as dynamic conformational ensembles [1–3]. They are abundant across all kingdoms of life [4,5] and implement a wide range of molecular functions [6–9]. These proteins/regions complement functional repertoire of ordered/structured proteins [10] and were shown to play important roles in several human diseases [11,12]. Studies of the intrinsically disordered proteins/regions improve our understanding of principles and mechanisms of protein folding and function.

Recent research in intrinsic disorder was stimulated by the availability of in-silico methods that predict disordered residues and regions in protein chains [13–15]. We focus on well-performing methods that are accessible to end users, either via web servers or standalone implementations. They include DISOPRED2 [16], IUPred [17], RONN [18], PROFbval [19], Norsnet [20], Ucon [21], PrDOS [22], DISOclust [23], MD [24], PreDisorder [25], POODLE [26], MFDp [27], PONDR-FIT [28], CSpritz [29], ESpritz [30], MetaDisorder [31], and SPINE-D [32]. These methods include publicly available versions of the best-performing disorder predictors from the 9th community-wide Critical Assessment of techniques for protein Structure Prediction (CASP9), such as PrDOS, DISOPRED, PreDisorder (also called MULTICOM), SPINE-D, POODLE, MFDp and DISOclust [33], and the top-performing predictors from CASP10 (based on our evaluation using publicly available results from the CASP10 site), such as PrDOS, DISOPRED, MFDp, POODLE, and PreDisorder. Although these methods provide accurate disorder predictions at the residue level [14,33], they make relatively substantial mistakes at the sequence-level, i.e., they usually over- or under-predict the overall amount of disorder in a given chain. A benchmark test of 10 recent predictors shows that the average mean absolute errors between the native and the predicted amount of disorder per chain vary between 15 and 39% [34]. In another benchmark of 19 predictors the average mean absolute errors ranged between 15 and 44% [14]. One explanation for these errors is that most of these methods, including the well-performing predictors in the recent CASPs such as DISOPRED2, MFDp, POODLE, PreDisorder, PrDOS, and SPINE-D, use a local/sliding sequence window to predict the disorder. We argue that information aggregated over the entire chain may reveal a sequence-level disorder bias [34]. Furthermore, these methods utilize multiple sequence alignment with PSI-BLAST, which impedes high-throughput analysis on a

* Corresponding author. Tel.: +1 780 492 5488; fax: +1 780 492 1811.
E-mail address: lkurgan@ece.ualberta.ca (L. Kurgan).

1570-9639/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.bbapap.2013.05.022

proteomic scale due to the relatively high computational cost. Our analysis reveals that a modern desktop computer requires approximately 350 s to calculate PSI-BLAST profile for a chain with about 400 amino acids (AAs). The calculation of these profiles over the human proteome with 70,000 proteins and the average chain size of 400 AAs would require over 280 days; a more accurate estimate is given in the Results and discussion section.

The sequence-level disorder content, defined as a fraction of disordered residues in a protein sequence (i.e., number of disordered residues divided by the total number of residues in a given chain), finds applications in many areas. It was used to estimate the abundance of intrinsic disorder in certain databases [35], protein families and classes [36–38], and complete proteomes [4,5,39]. The content was also utilized in the analysis of intrinsic disorder-related protein functions [40–42]. Varying amounts of disorder content values were reported for proteins associated with different diseases [11,12,43]. Furthermore, the predicted disorder finds more "practical" applications in functional proteomics [10], with examples in target selection in structural genomics [44–47] and prediction of functional sites [48]. However, to date only one method, DisCon [34], was designed to accurately predict the disorder content and this method utilizes PSI-BLAST.

With rapid advancements and decreasing costs of high-throughput sequencing technologies, we anticipate a growing need to provide time-efficient analysis of the disorder content. To this end, we aim to provide a fast and accurate method to predict the disorder content in a given protein chain. This is motivated by the fact that the existing and accurate disorder predictors are relatively slow, that the quality of the disorder content calculated from their predictions requires further improvements, and that the existing disorder content predictor DisCon is also time-inefficient. The three main advantages of our support vector Regression-based Accurate Predictor of Intrinsic Disorder (RAPID) are:

- Speed; we use fast-to-compute inputs and prediction model, which allows predicting an entire eukaryotic proteome in 1 h or less on a modern desktop computer.
- Sophisticated design; we hand-crafted and selected inputs based on information extracted from predicted per-residue disorder, sequence complexity, and selected physicochemical properties of AAs that are aggregated over the input chain.
- High-quality predictions; tests on 2 diverse benchmark sets show that RAPID compares favorably against DisCon and a comprehensive set of state-of-the-art disorder predictors.

We also applied RAPID to analyze disorder in 200+ eukaryotic proteomes, with a more detailed analysis for the human proteome.

2. Materials and methods

2.1. Datasets and evaluation protocols

RAPID was designed and tested on the MxD dataset, which was originally developed in [27] and used to design and validate DisCon [34]. This dataset contains 514 proteins with pairwise sequence identity <25% and with disorder annotation that were extracted from protein data bank (PDB) [49] and DisProt [50] using procedures described in [33] and [51]. This dataset was split at random into two equally-sized sets of chains. One set of 257 chains constitutes the TRAINING dataset. The entire design, which includes selection of input features and parameterization of the prediction model, was performed utilizing 5-fold cross validation on the TRAINING dataset. The other set of 257 chains was further expanded to include recent depositions from DisProt and PDB to form a relatively large, new TEST dataset. We considered chains added to DisProt after release 4.6 (which was used to build the MxD dataset) and to PDB after Aug. 1, 2011. Among these chains we removed proteins that share >25% sequence identity to any chain in the MxD dataset and the training datasets used by one of the most recent disorder predictors CSpritz [29]. The remaining 104 proteins were annotated the same way as the chains in the MxD dataset. The resulting new TEST set has 257 + 104 = 361 chains that share low (<25%) identity with the proteins in the TRAINING set. The TRAINING and TEST datasets are available at http://biomine.ece.ualberta.ca/RAPID/. We also use 95 chains from the most recent CASP10 experiment, for which chains and disorder annotations were downloaded from http://predictioncenter.org/download_area/CASP10/. We collected disorder predictions for these chains from all participating predictors in CASP10, which are available at the same URL, to compare with RAPID.

To evaluate predictive performance of RAPID, the model built on the TRAINING dataset was tested on the new TEST and CASP10 datasets and compared against state-of-the-art in the field. Fig. 1 shows that these test datasets have substantially different distributions of the disorder content values. The TEST dataset has more proteins with larger content values including a relatively large fraction of fully disordered proteins (with content = 1), while the CASP10 set includes a large fraction of proteins with low amounts of disorder and fully structured proteins (with content = 0). To compare, there are 27% and 9% of proteins with over 0.25 disorder content in the TEST and CASP10 datasets, respectively.

2.2. Evaluation criteria

The predictions were evaluated using the same criteria as used in [34], including:

$$\text{Mean Absolute Error (MAE)} = \sum_{i=1,\ldots n} \frac{|x_i - y_i|}{n}$$

$$\text{Mean Squared Error (MSE)} = \sum_{i=1,\ldots n} \frac{(x_i - y_i)^2}{n}$$

$$\text{Pearson Correlation Coefficient (PCC)} = \sum_{i=1,\ldots n} \frac{(x_i - x_m)(y_i - y_m)}{(n-1)s_x s_y}$$

where n is the number of protein chains in the dataset; $x^i \in X$ is the predicted disorder content and $y^i \in Y$ is the native disorder content for the ith ($i = 1,2,\ldots n$) protein chain; x^m and y^m are the mean values of populations X and Y; and s^x and s^y are the standard deviations of X and Y.

We evaluated the statistical significance of the differences between the content predictions of RAPID and each of the other considered predictors. For each test dataset we randomly selected 70% of proteins (to have large enough sample for the CASP10 dataset that has 95 chains) to calculate the corresponding MAE, MSE and PCC values. This is repeated 10 times and we compared the corresponding 10 paired results for each of the three measures. Given that the measurements are normal, as tested with the Anderson–Darling test at 0.05 significance, we utilized the paired t-test to investigate significance; otherwise we used the Wilcoxon test. Differences between were assumed statistically significant when p-value < 0.05.

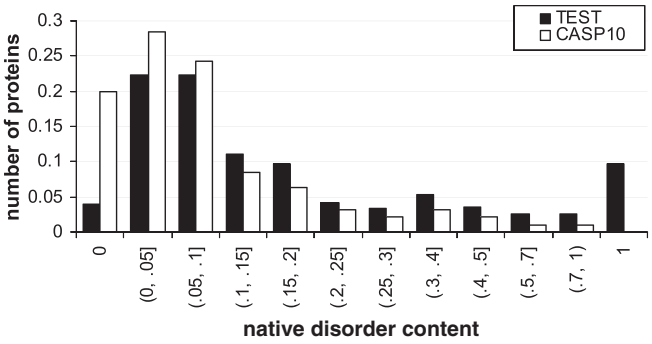

Fig. 1. Distribution of fraction of proteins (y-axis) in given intervals of the native disorder content for the TEST and CASP10 datasets. The x-axis shows the content binned to 0.05 wide intervals including values of 0 (fully structured proteins) and 1 (fully disordered proteins) on both ends.

Moreover, we evaluated prediction of proteins with large disorder content based on the content predictions. We binarized the native and predicted content to classify each protein as either having "large" amount of disorder (≥25%) or "small" amount of disorder (<25%), and we computed Mathews correlation coefficient (MCC) between these native and predicted binary labels over a given test dataset:

$$MCC_{.25} = (TP * TN - FP * FN) / \sqrt{[(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)]}$$

where TP is the number of true positives (correctly predicted proteins with large amount of disorder), FP denotes false positives (proteins with small disorder content that were predicted as having large content), TN denotes true negatives (correctly predicted proteins with small amount of disorder), and FN stands for false negatives (proteins with large disorder content that were predicted as having small content). The MCC values range between −1 and 1 and they are equal to zero when all proteins are predicted as positives or negatives. The use of MCC is motivated by the fact that our datasets are imbalanced.

2.3. RAPID predictor

RAPID generates predictions using four steps, see Fig. 2. In the first step, the low/high complexity regions, per-residue disorder predictions, and various physicochemical and biochemical properties are computed for the input protein using SEG program [52], IUPred [17], and the AAindex database [53], respectively. Next, this information is encoded using 7 custom designed features, which are empirically selected using TRAINING dataset from a large pool of 3000+ features. These 7 features are inputted into a support vector regression (SVR) model that predicts the disorder content. Selection of SVR was motivated by empirical results that show that it outperforms an alternative ridge regression model (see below) and the fact that it generates predictions quickly. Lastly, the predicted content values are set to the unit interval, i.e., negative values are set to 0, and predictions > 1 are set to 1.

Our goal is to provide a method that rapidly predicts the disorder content and this motivated the choice of the prediction model and the features/information that are computed from the input sequence. IUPred was selected to provide per-residue disorder predictions based on its strong predictive performance in a recent comparative review [14], where its average mean absolute error outperformed predictions of 18 other disorder predictors, and the fact that it is fast easy to predict as it does not utilize alignment. Similarly, the fast-to-compute SEG algorithm was used to annotate the low and high complexity regions. We derived four groups of features from a given protein chain and physicochemical and structural properties of its amino acids (AAs), which aggregate information over the input chain. These feature sets are defined in detail in the Supplement (available at http://biomine.ece.ualberta.ca/RAPID/Supplement.pdf). Total of 3758 features, which comprehensively cover possible numerical inputs to the prediction model that can be extracted from the above information, were generated.

Some of these features are likely irrelevant to the prediction and also could be redundant/correlated with each other. Thus, we performed two-step feature selection to collect a small subset of relevant and non-redundant features. First, we removed the irrelevant and redundant features using a coarse-grained evaluation based on correlation with the native disorder content, and next we executed a wrapper-based feature selection using the remaining features. The second step also includes computation and parameterization of the prediction models. These two steps were performed utilizing cross-validation on the TRAINING dataset.

In the first step, we performed the correlation-based feature selection where we calculated the PCC values (average on the 5 training folds based on the 5-fold cross validation on the TRAINING dataset) with the disorder content and removed irrelevant features that had absolute PCC < 0.16; this cut-off corresponds to a visible peak in the histogram of the absolute PCC values shown in Fig. S1 in the Supplement. Next, we removed redundant features, i.e., features that are highly correlated with each other. Specifically, we sorted features according to their absolute PCC values and starting with the feature that had the highest PCC value we added a subsequent feature from the sorted list only if its PCC value (averaged over the 5 training folds) with all the previously chosen features is <0.7. The features that were not added are highly correlated (PCC ≥ 0.7) with the selected features, and thus were removed. This step resulted in the selection of 284 features.

In the second step, we used the wrapper-based sequential forward selection on the features obtained from the first step. We considered two predictors, linear ridge regression and SVR with Radial Basis Function kernel, to implement the wrapper. First, these predictors were parameterized using top 5% (5% * 284 = 14) of the 284 features (selected based on the average PCC values computed in the first step); next they were used to perform the wrapper-based selection; and finally they were parameterized again using the selected feature set. These parameterizations and selections were performed using TRAINING dataset and a variant of the 5-fold cross validation called 4 + 1-fold cross validation. This protocol includes 4-fold cross validation on 4 out of the 5 original folds and additional test in which these four folds are used together to build a model that is tested on the set-aside 5th fold. The use of the 4 + 1-fold cross validation helps to reduce overfitting into the training dataset and was successfully used in our prior studies [48]. The parameterization was run utilizing grid search over the ridge parameter r for the ridge regression, and complexity parameter C and kernel parameter γ for the SVR model. We considered $r = 10^i$ where $i = -10, -9, \ldots, 5$, and $C = 2^n$ and $\gamma = 2^n$ where $n = -5, -4, \ldots, 5$. The parameter set with the lowest MAE value on the 4 + 1-fold cross validation on TRAINING dataset was selected. The results from the initial/first parameterization, which are shown in Table S1 in the Supplement, reveal that ridge regression is not sensitive to the ridge value (the same results were obtained for initial/default and final/optimized value of r), while parameterization helps the SVR model to reduce MAE by about 10%. Next, the two parameterized predictors were used to perform sequential forward feature selection from the set of 284 features. First, we computed the MAE value for each of the features used individually to predict the native disorder content based on the 4 + 1-fold cross validation on the TRAINING dataset. We sort the 284 features based on these MAE values and select the best performing feature. We accept the next ranked feature into the current list of selected features if the addition

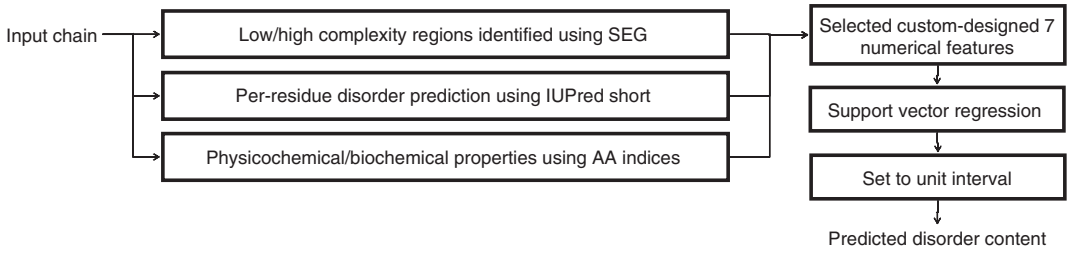

Fig. 2. Architecture of the RAPID predictor.

of this feature decreases the MAE value by at least 0.001 on the 4 + 1-fold cross validation when compared with the set before the addition. We scan the sorted list of features once. Fig. S2 in the Supplement shows the progress of the selection process where both prediction models saturated the improvements in MAE after selecting 7 features. Both predictors are again parameterized using their selected 7 features and the same approach as for the initial parameterization. The results are summarized in Table S1 in the Supplement. This second/final parameterization does not provide further improvements. The MAEs obtained using the parameter values selected through the initial parameterization are virtually identical to the results with the re-optimized parameters. The SVR obtains substantially lower MAE on the 4 + 1-fold cross validation on the TRAINING dataset compared to the ridge regression, i.e., 0.151 vs. 0.160. Consequently, the SVR-based model was implemented in the RAPID predictor.

RAPID utilizes 7 selected custom-designed features, which are summarized in Table S2 in the Supplement. They use information from per-residue predictions with IUPred, which agrees with the strong predictive performance of IUPred shown in [14]; AA index that quantifies B-factors, which concurs with the observations in [34,54]; AA index that quantifies hydrophobicity, which is consistent with the observations in [55,56]; AA index that represents propensity for helical conformations, which was discussed in [57,58]; AA index related to side chain interactions, which were pointed out to be relevant to disorder in [6]; and finally they make use of the low sequence complexity regions, which agrees with [59]. Although these factors are already known to be important for characterization and prediction of disorder, they are quantified using custom-designed numerical features and efficiently combined through empirical feature selection. These two latter aspects allow RAPID to provide strong predictive performance. Table S2 in the Supplement reveals that the average absolute MAEs of these features, when used individually to predict disordered content on the TRAINING dataset, are substantially larger than the MAE of their combinations that is implemented in RAPID, i.e., the best MAE of an individual feature equals 0.181, see Table S2 in the Supplement, compared to 0.151 of RAPID, see Table S1 in the Supplement. This suggests that disorder is a complex phenomenon that can be accurately predicted by combining multiple complementary information sources, which is in agreement with prior studies [24,27,34]. We visualize relationships between the top three (according to MAE in Table S2 in the Supplement) features and the disorder content in the Fig. 3. The points correspond to proteins in the TRAINING dataset, which are color-coded according to their native disorder content, from blue for low disorder content to red for the high content. The 3 features (shown on the 3 axes) have positive correlations with the native disorder content, which means that higher averaged per chain predicted real-value propensity for disorder (IUPred-p-avg axis), higher average length of predicted disordered segments (IUPred-b-avgDisL axis) and higher estimated per-chain flexibility (VINM940104_avg axis) are associated with larger disorder content. Most importantly, combination of these complementary features improves prediction of disorder content, i.e., the points are relatively well separated by colors/content in this 3-dimentional space, while projection of these points on just one axis (say, VINM940104_avg axis) would mix proteins of varying colors/content together.

3. Results and discussion

3.1. Comparison with existing residue-level and disorder content predictors

RAPID is compared with DisCon and a representative set of 21 modern residue-level disorder predictors on the TEST dataset. The residue-level predictors include DISOPRED2, IUPred in two versions, short and long, PROFbval, NORSnet, Ucon, PrDOS, DISOclust, MD, PreDisorder, MFDp, PONDR-FIT, CSpritz in two versions: short and long, ESpritz in 6 versions including models optimized for low false positive rate (FPR) and high S_w using NMR-based annotation of disorder (ESpritz NMR-FPR and ESpritz NMR-S_w), DisProt-based annotations of disorder (ESpritz DP-FPR and ESpritz DP-S_w), and X-ray crystals-based annotations (ESpritz Cx-FPR and ESpritz Cx-S_w) and SPINE-D. Table 1 reports MAE, MSE, and PCC values between the native and the predicted disorder content and $MCC_{.25}$ values for the predictions of proteins with large amount of disorder for these 23 methods. RAPID provides the lowest MAE and MSE, the highest PCC, and the fourth best $MCC_{.25}$. The second best PONDR-FIT short has MSE worse by (0.054–0.049)/0.049 = 10%, MAE worse by (0.154–0.141)/0.141 = 9%, and a slightly

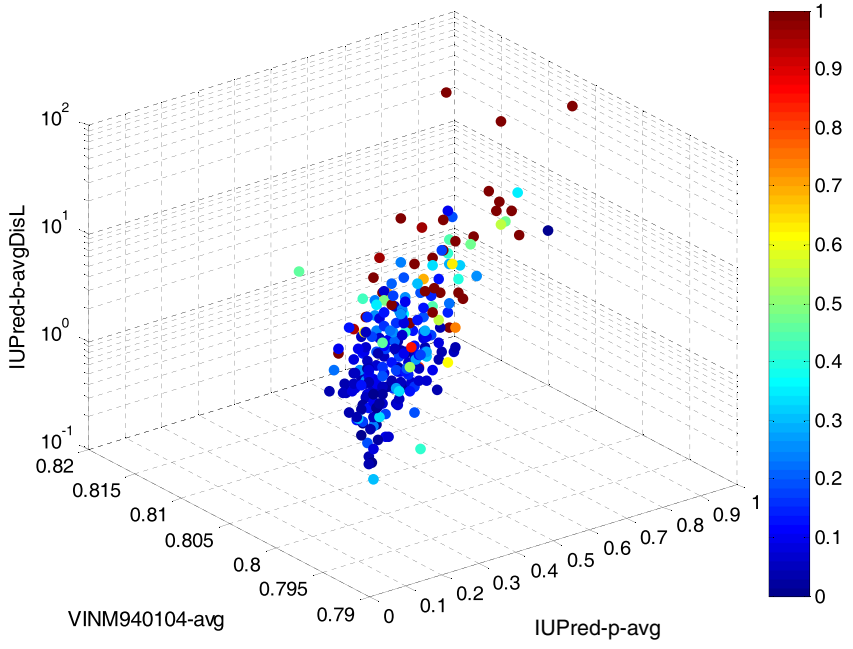

Fig. 3. Scatter plot of the top three features (according to MAE values shown in Table S2 in the Supplement) used by RAPID. Color denotes the native disorder content of a given point (defined using a scale on the right) that represents a protein from the TRAINING dataset; dark blue stands for fully structured proteins and dark red for fully disordered proteins.

Table 1
Comparison of RAPID, DisCon and 21 modern disorder predictors on the TEST dataset. The methods are divided into single-sequence methods that do not utilize PSI-BLAST profiles, and methods that use the profiles. The best values are shown in bold font, and best values for (fast) methods that do not use PSI-BLAST are given in italics. Methods are sorted in the ascending order by the MSE values within each sub-group. +/−/= indicates that RAPID is statistically significantly better/worse/not different at p-value < 0.05 than another method. PreDisorder and PrDOS could not predict 8 proteins and they were evaluated on the remaining 353 chains.

Predictor		MAE		MSE		PCC		$MCC_{0.25}$
Type	Name	Value	Sig	Value	Sig	Value	Sig	
Single-sequence methods that do not utilize PSI-BLAST profiles	RAPID	**0.141**		**0.049**		**0.684**		*0.483*
	IUPred short	0.146	+	0.055	+	0.637	+	0.425
	ESpritz Cx-FPR	0.145	+	0.060	+	0.608	+	0.421
	ESpritz Cx-S_W	0.175	+	0.062	+	0.624	+	0.401
	IUPred long	0.157	+	0.067	+	0.590	+	0.423
	ESpritz NMR-FPR	0.178	+	0.068	+	0.553	+	0.434
	ESpritz DP-FPR	0.173	+	0.084	+	0.601	+	0.418
	ESpritz NMR-S_W	0.252	+	0.101	+	0.537	+	0.243
	ESpritz DP-S_W	0.226	+	0.133	+	0.554	+	0.462
Methods that utilize PSI-BLAST profiles	PONDR-FIT	0.154	+	0.054	+	0.669	+	0.494
	DisCon	0.166	+	0.054	+	0.642	+	0.413
	Ucon	0.168	+	0.060	+	0.619	+	**0.511**
	PrDOS	0.146	+	0.061	+	0.582	+	0.486
	DISOPRED2	0.159	+	0.064	+	0.599	+	0.472
	PreDisorder	0.208	+	0.075	+	0.611	+	0.376
	MFDp	0.175	+	0.077	+	0.623	+	0.464
	SPINE-D	0.203	+	0.079	+	0.612	+	0.382
	CSpritz short	0.206	+	0.079	+	0.588	+	0.351
	CSpritz long	0.194	+	0.092	+	0.609	+	0.428
	NORSnet	0.185	+	0.095	+	0.447	+	0.392
	MD	0.199	+	0.096	+	0.610	+	0.413
	DISOclust	0.242	+	0.099	+	0.549	+	0.261
	PROFbval	0.402	+	0.195	+	0.372	+	0.102

lower PCC by (0.684–0.669)/0.669 = 2%. RAPID also offers relatively good $MCC_{.25}$, which is slightly lower than the best value obtained by UCon (0.483 vs. 0.511) and higher than for any method that does not utilize PSI-BLAST (0.483 vs. 0.462). Moreover, RAPID outperforms the other disorder content predictor DisCon by a relatively wide margin, in spite of the fact that DisCon uses PSI-BLAST profiles. We note that the results for DisCon are slightly different than the results reported in [34], i.e., MAE of 0.166 vs. 0.156; MSE of 0.054 vs. 0.050, and PCC of 0.64 vs. 0.68, since here we use a different, substantially larger test dataset. The improvements offered by RAPID over the other considered methods on the TEST dataset are shown to be statistically significant. The magnitude of these improvements ranges between 10% (compared to PONDR-FIT) and 171% (ESpritz DP-S_W) for MSE, between 3.5% (IUPred short) and 78% (ESpritz NMR-S_W) for MAE, and between 2.2% (PONDR-FIT) and 53% (NORSnet) for PCC; we exclude PROFbval that is primarily used to predict B-factors and is historically included in the comparative evaluations of disorder predictors.

We also compare RAPID against the state-of-the-art in disorder prediction based on the results of the CASP10 experiment. We evaluated the publicly available results, which we downloaded from http://predictioncenter.org/download_area/CASP10/ in mid December 2012, for all groups that submitted predictions for all 95 targets that were assessed under the disorder prediction category. The predictive performance of these methods including MAE, MSE, PCC, $MCC_{0.25}$ and AUC, which evaluate residue-level prediction, is summarized in Table 2. For each group we list their "closest" publicly available disorder predictor, however the predictions are often performed using customized in-house implementations. RAPID has secured second best MSE at 0.014, second best $MCC_{0.25}$ at 0.509, and fourth best PCC at 0.48 on this benchmark dataset. The differences in MSE are statistically significant, except for the groups 288, 369, and 478 that have comparable predictive quality. For PCC, RAPID significantly outperforms 12 participants and provides correlations that are not significantly different compared with the 7 remaining groups. For MAE, 4 groups significantly outperform RAPID, while the remaining 15 groups offer predictions with worse scores. Although our method did not outperform some of the CASP10 participants, most of these methods utilize PSI-BLAST (including all methods that outperformed RAPID) while our predictor offers the additional benefit of providing fast predictions that can be applied on a proteomic scale.

3.2. Comparison of runtime

Runtime is a key factor that determines whether a given predictor can be applied in a high-throughput manner. The existing disorder

Table 2
Comparison between RAPID and groups participating in CASP10 that predicted all 95 targets considered under the disorder prediction evaluation. The methods are divided into single-sequence methods that do not utilize PSI-BLAST profiles, methods that use the profiles, and methods where this information is unknown. [Public predictor] denotes the closest publicly available/published method developed by the same group. Best values are shown in bold font. Predictors are sorted in the ascending order by the MSE values within each sub-group. +/−/= indicates that RAPID is statistically significantly better/worse/ not different with p-value < 0.05 than another method. The AUC values are computed for the per-residue predictions and they are not available for RAPID.

Predictor		MAE		MSE		PCC		$MCC_{.25}$	AUC
Type	Group number [public predictor]	Value	Sig	Value	Sig	Value	Sig		
Single-sequence Methods that do not utilize PSI-BLAST profiles	RAPID	0.081		0.014		0.476		0.509	n/a
	327 [ESpritz]	0.099	+	0.025	+	0.407	+	0.254	0.850
	380 [Espritz]	0.143	+	0.038	+	0.293	+	0.133	0.846
Methods that utilize PSI-BLAST profiles	288 [MFDp]	**0.062**	−	**0.013**	=	0.497	=	0.246	0.882
	369 [PrDOS]	0.067	−	0.014	=	0.436	=	0.407	**0.896**
	478 [MFDp]	0.067	−	0.014	=	0.428	=	0.085	0.885
	170 [DISOPRED]	0.073	−	0.017	+	0.293	+	−0.047	0.880
	84 [MFDp]	0.104	+	0.021	+	0.200	+	0.141	0.822
	222 [PreDisorder]	0.113	+	0.027	+	0.508	=	0.504	0.864
	216 [POODLE]	0.119	+	0.028	+	0.294	+	0.198	0.866
	424 [PreDisorder]	0.132	+	0.028	+	**0.519**	=	**0.513**	0.848
	413 [SPINE-D]	0.156	+	0.040	+	0.386	+	0.098	0.859
	125 [PreDisorder]	0.160	+	0.047	+	0.453	=	0.388	0.839
	496 [MetaDisorder]	0.145	+	0.051	+	0.350	+	0.362	0.801
	484 [CSpritz]	0.168	+	0.052	+	0.356	+	0.167	0.822
	183	0.233	+	0.077	+	0.382	=	0.163	0.772
	494 [MetaDisorder]	0.295	+	0.114	+	0.323	+	0.037	0.772
	273 [DISOclust]	0.246	+	0.119	+	0.128	+	0.183	0.819
Unknown	180	0.102	+	0.020	+	0.349	+	0.156	0.861
	167	0.146	+	0.034	+	0.156	+	0.198	0.592

predictors can be divided into two groups: methods that utilize PSI-BLAST to derive evolutionary profiles, and the single-sequence methods that do not. The main computational cost of methods in the first group, which includes the only existing disorder content predictor DisCon and top 6 (based on AUC) methods in CASP10, is the multiple sequence alignment, and thus we approximate their runtime by the time to run PSI-BLAST. Fig. 4 compares average runtime of PSI-BLAST and the methods from the second group (IUPred, Espritz, and RAPID) based on predictions on the TEST dataset. We sorted chains based on their length and divided them into 10 equally-sized subsets with increasing protein size. The averaged runtime is plotted against the averaged protein size over these subsets. Although the absolute runtime values depend on a computer hardware used, we focus on relative differences which are hardware independent. PSI-BLAST takes a considerable amount of time to run, which increases by an order of magnitude between short and very long chains. ESpritz, which is 3 orders of magnitude faster than PSI_BLAST, is also characterized by an increase in the runtime by an order of magnitude with the increase of the chain size. The fastest IUPred and RAPID have 3–4 orders of magnitude lower runtime compared to PSI-BLAST and their execution time does not increase for longer chains. The average, over all chains, runtimes for IUPred, RAPID, ESpritz and PSI-BLAST are 0.05, 0.06, 0.4, and 347 s, respectively. Although RAPID takes longer to compute than IUPred, the difference is relatively small while RAPID provides improved predictive performance.

We used linear approximations from Fig. 4, which have good fit into the measured data, to estimate runtime of IUPred, RAPID, Espritz and PSI-BLAST on the entire human proteome. We compare these approximations with the measured time of the fast IUPred and RAPID, see Table 3. The error in the runtime estimate is relatively small, up to 25% higher than the measured value. Assuming similar errors for ESpritz and PSI-BLAST, the results demonstrate substantial advantage offered by RAPID and IUPred compared to the other predictors. Using Espritz and any method that applies PSI_BLAST it takes 5.5 and 5043 h, respectively, to calculate predictions for the human proteome, compared to about 1 h for IUPred and RAPID. Our results are consistent with the results on a 1% of human proteome from [30], where ESpritz was shown to be 3–4 orders of magnitude faster than PSI-BLAST and an order of magnitude slower than IUPred.

3.3. Disorder in eukaryotic proteomes

Using RAPID, we predicted disorder content for 200+ fully sequenced eukaryotic proteomes (~3.2 million chains) collected from the Uniprot database [60] in July 2012; see Table S3 in the Supplement (available at http://biomine.ece.ualberta.ca/RAPID/Supplement.pdf). We also estimated structural coverage of each proteome following the protocol from [61]. For every chain, we run 3 rounds of PSI-BLAST against chains in the PDB database. A given protein is considered "structured" if it has a hit in PDB with E-value < 0.001 that has ≥50 AAs in length.

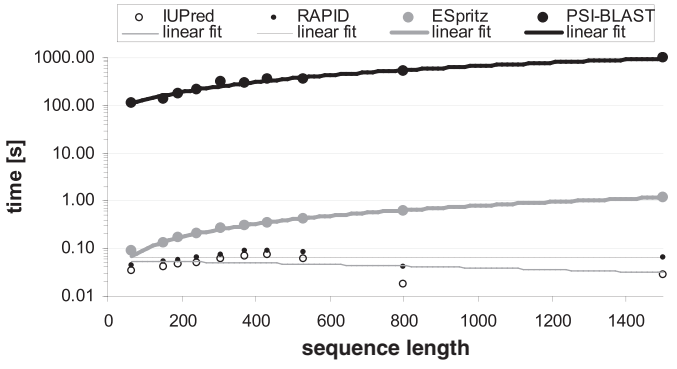

Fig. 4. Runtime (*y*-axis in logarithmic scale) in the function of sequence size (*x*-axis). The measurements were made using a modern desktop computer using the TEST dataset. The runtime was averaged for chain sizes binned into 10 intervals that include similar numbers of proteins.

Table 3
Actual and estimated runtime in minutes of IUPred, RAPID, ESpritz, and PSI-BLAST for the disorder prediction on the human proteome using a modern desktop computer.

	IUPred	RAPID	ESpritz	PSI-BLAST
Actual runtime [min]	34.2	63.9	–	–
Estimated runtime [min]	47.2	75.4	331.4	302,625

Our study complements and expands previous works, including studies that analyzed disordered binding regions in 44 eukaryotic proteomes [62], relation between disorder, proteome size, and organism complexity in 53 eukaryotes [63], abundance of disorder in 67 eukaryotes [5], and the largest to date analysis of 194 eukaryotes for which disorder was contrasted against prokaryotes [64]. We investigate other interesting aspects including relation of disorder with structural coverage and chain length, and we characterize abundance of chains with large amount of disorder and fully disordered chains across various eukaryotic phyla. Moreover, we analyze a large number of species; only Pancsa and Tompa [64] considered comparable number of proteomes, but they used a slightly less accurate IUPred to perform the disorder predictions.

Fig. 3 shows relations between the disorder content aggregated over entire proteomes and the corresponding structural coverage. We separate proteomes by their kingdoms/phyla into *Alveolata*, *Fungi*, *Metazoa*, *Viridiplantae*, and others that have too few species to be grouped. Fig. 5A shows that structural coverage is lower for proteomes with larger average disorder content. This agrees with the observation that chains with disordered segments are harder to

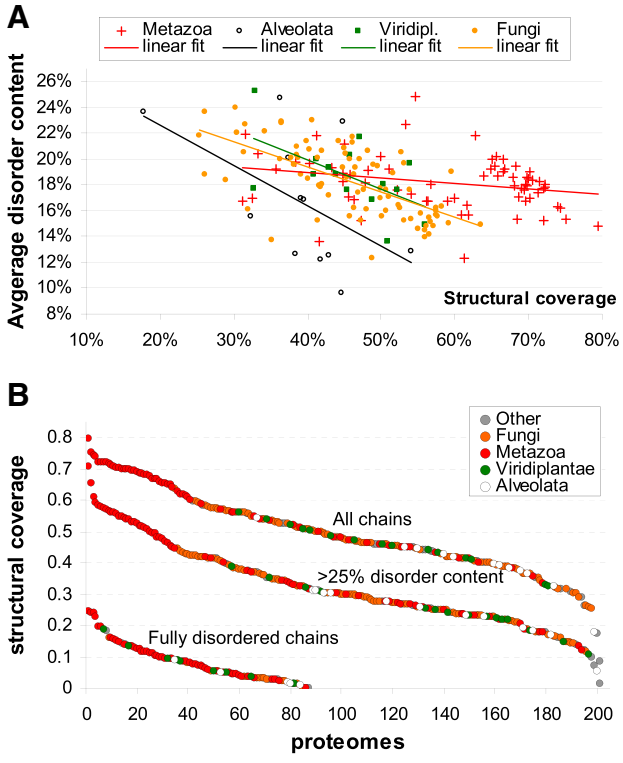

Fig. 5. Scatter plot of relation between predicted disorder content and structural coverage (panel A) and structural coverage for all chains, chains with at least 0.25 disorder content and fully disordered chains (panel B). In Panel B we plot only the protein sets that have over 50 chains. Eukaryotic proteomes are divided by their kingdoms/phyla into *Alveolata*, *Fungi*, *Metazoa*, and *Viridiplantae*.

solve using the dominant structure determination approach via X-ray crystallography [45]. However, this trend is weaker for animals where the coverage reaches 70–80% in spite of their average disorder content ranging between 15 and 20%; this predicts that many eukaryotes in other phyla can be also brought to these levels of coverage. *Homo sapiens* has lower coverage at 55% coupled with relatively high disorder content at 25%. Fig. 5B shows coverage for whole proteomes (top line) and subsets of proteomes with proteins that have ≥25% of disorder (middle line) and only fully disordered chains (bottom line); we show subsets that have over 50 chains. As expected, coverage drops substantially for chains with large amount of disorder, on average by 17% when considering chains with ≥25% disorder content. However, even some fully disordered chains have solved structures. Interestingly, over 20% of fully disordered chains in several animal species have structures. This usually concerns fragments of their chains, such as individual domains or interaction sites, for which the structure could be stabilized, e.g., through an interaction with another molecule [47].

Fig. S3A in the Supplement shows strong correlation between the disorder content and fraction of chains with large (>25%) disorder content. Some eukaryotes including parasites (*Kinetoplastida Duttonella* and 3 species of *Apicomplexa*), plants (*Streptophyta Oryza*), fungi (8 species of *Dikarya*), and animals (*Homo sapiens*) have over 1/3 of chains with the large amount of disorder. *Homo sapiens* has 38% of proteins with the large disorder content, which is a higher value compared to the previous estimate of ~22% obtained on a smaller subset of human proteome [4]. However, the fraction of fully disordered chains has weaker (compared to the fraction of chains with large disorder content) correlation with the overall disorder content, see Fig. 6. In particular, animal species have larger fractions of fully disordered chains relative to other phyla. Overall, close to 90 out of 201 proteomes, particularly in animal organisms, have over 50 fully disordered chains, see bottom line in Fig. 5B. Similar to [64], we observe that larger proteomes are characterized by larger disorder content, except for plants where this trend disappears and animals where the correlation is weaker; see Fig. S3B in the Supplement.

We also analyze disorder content values in the context of the corresponding protein counts and chain length for individual eukaryotic proteomes grouped by kingdoms/phyla. Fig. 7A shows color-coded fractions (ranging from low fractions in white to high fractions in dark red) of chains in a given proteome (each horizontal line corresponds to one proteome) over varying amounts of content that are listed on the horizontal axis. It reveals that majority of proteins have relatively low amounts of disorder, i.e., darker red is concentrated on the left. However, the entire range of disorder content is colored (i.e., non-white colors are distributed over the entire range of disorder content), which means that virtually all proteomes include chains with significant amounts of disorder. The right-most column in Fig. 7A (i.e., for the disorder content equal 1) demonstrates that numbers of fully disordered chains are fairly substantial; they are as high as up to 3.4% of a given proteome and we found 16 proteomes with ≥1% of fully disordered chains, particularly in animals and plants. The white spots on the right side that appear in a large number of fungi, protists, and "other" species, reveal that they do not have proteins with large disorder content. Fig. 7B shows predicted color-coded disorder content (ranging from content at 0% in white to 100% in dark red) in the function of chain length that is listed on the horizontal axis; the right-most column aggregates all chains longer than 3000 AAs. The figure demonstrates that large disorder content is primarily found in relatively short proteins, with a length of about 100 or fewer AAs. However, we observe a second peak of the high disorder content values (cluster of dark red points) for the chain lengths of about 1500 and 1000 for many animal and fungal species, respectively. Interestingly, the disorder in these long chains with large disorder content plays different functional roles compared with the short disordered chains; we recently investigated that using a smaller subset of eukaryotes [65].

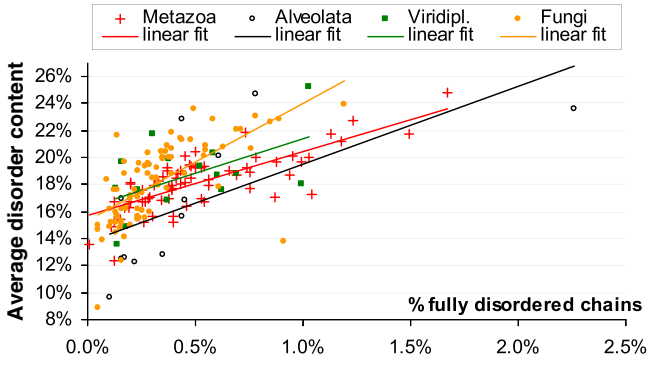

Fig. 6. Scatter plot of relation between predicted disorder content and fraction of fully disordered chains for the eukaryotic proteomes. Eukaryotic proteomes are divided by their kingdoms/phyla into *Alveolata*, *Fungi*, *Metazoa*, and *Viridiplantae*.

3.4. Functional analysis of disorder in Homo sapiens

Functional analysis of disorder in eukaryotes was previously done on a smaller scale, for yeast in [4] and using 28,057 human proteins annotated with GO terms in [42]. We perform a more comprehensive study using predictions from RAPID for 56,392 annotated proteins from the human proteome. We collected gene ontology annotations [66] for 3 major categories including cellular component (CC), molecular function (MF) and biological process (BP). We assume that GO annotations that occur in at least 100 chains have sufficient statistical power to calculate enriched in disorder. Similar to an earlier analysis on yeast [4], we compare disorder content in a set of chains with a given annotation to the content in a randomly selected, from the entire human proteome, set of the same number of chains. We also assure that the randomly selected chains have similar length (±10%); this accommodates for chain-size bias of the disorder content per Fig. 7B. This is repeated 10 times, each time selecting 50% of GO annotated chains, and we evaluate the significance of the differences in the disorder content between these two vectors. For normal measurements (as tested with the Anderson–Darling test) we applied paired t-test; otherwise we used the Wilcoxon test. The enrichment (increase in the disorder content) is assumed significant if p-value < 0.05 and the difference between the averaged (over 10 repetitions) disorder content in the annotated and randomly selected sets of chains is greater than 0.1.

Total of 188 biological processes, 75 molecular functions, and 83 cellular components were considered, out of which 40 (21%), 16 (21%), and 22 (27%), respectively, were found to be enriched in disorder. These significantly enriched annotations are shown in Fig. 8. We found numerous annotations that point to the substantial role of disorder in various RNA- and DNA-related functions and numerous key cell processes including mitosis, differentiation, morphogenesis, arrest, etc. Several annotations point to the strong enrichment of disorder in the nucleosome, which agrees with [38], transcription factor complex, centrosome, nucleus, and several other cellular compartments. Some of the terms that we found to be enriched have been identified in the previous (more limited) studies including transcription factor complex, nucleolus, nucleus and actin cytoskeleton from the CC category [4]; DNA binding, nucleic acid binding, kinase and RNA binding in the MF category [4,42]; and transcription from RNA pol promoter, regulation of transcription from RNA pol promoter, transcription (DNA dependent), positive regulation of transcription (DNA dependent), rRNA processing, cell cycle and DNA replication from the BP category [4,42]. These correlated findings provide validation of our results. Overall, only 8 out of 40 processes, 4 out of 16 functions, and 4 out of 22 cellular components that we found to be enriched were also listed in the other two studies. The numerous other annotations provide

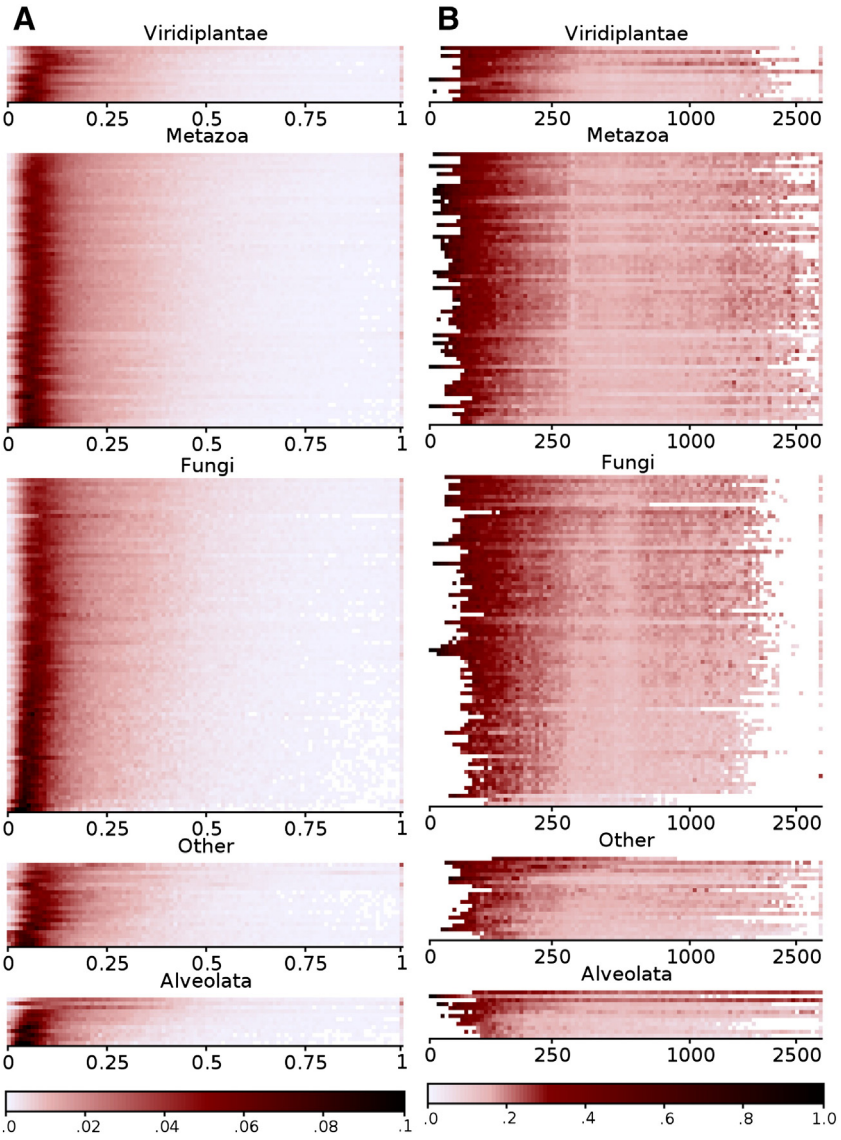

Fig. 7. Disorder content values in the context of protein counts (Panel A) and chain length (Panel B). Each horizontal line corresponds to one proteome; these proteomes are grouped together by their kingdoms/phyla and are sorted within each kingdom by their average content. Panel A shows distributions of the number (fraction) of proteins for a given value of the disorder content. The disorder content values, which range between 0 (structured protein) and 1 (fully disordered protein), are given on the horizontal axis. The fractions of chains for a given small range of values of content are denoted by colors, where white denotes 0% of chains and dark red denotes 10% or more of chains in a given proteome; see the color scale at the bottom of the panel. Panel B shows distributions of the disorder content values for a given protein chain size. The chain sizes, which range between 0 and 3000 amino acids, are given on the horizontal axis. The disorder content values (calculated as an average over chains in a given small interval of chain sizes) are denoted by colors, where white is for fully structured chains (0% of disorder content) and dark red for fully disordered chains (100% of disorder content); see the color scale at the bottom of the panel. White cells also denote that a given chain size interval has <10 chains, which means that the average content could not be reliably calculated. The right-most column provides the disorder content values for proteins with over 3000 amino acids.

new insights into functional roles and cellular localization of disorder in *Homo sapiens*.

4. Conclusions

We developed an accurate and fast predictor of disorder content called RAPID. Our custom-designed solution utilizes multiple complementary information sources that are combined based on empirical feature selection and an efficient SVR model. Empirical tests reveal that RAPID provides competitive predictive quality when compared to a comprehensive set of state-of-the-art disorder predictors while it is also very fast to compute, i.e., it predicts an average-sized eukaryotic proteome in <1 h on a modern desktop computer. RAPID is available at http://biomine.ece.ualberta.ca/RAPID/ as a web server for batch predictions of up to 75,000 chains (an entire proteome). This URL includes the Supplement and the TRAINING, TEST, and the human proteome datasets with the disorder content annotations.

RAPID was used to perform large-scale characterization of disorder in 200+ eukaryotic proteomes. We show several interesting observations concerning relations between disorder and structural coverage. Structural coverage is lower for proteomes with larger average disorder content, but this (expected) trend is weaker for animals where coverage reaches ~70% in spite of the fact that their disorder content is fairly high. We show that structural coverage is substantially lower for proteins with large amounts of disorder; however, even some fully disordered proteins have solved structures, particularly in animal organisms.

We also characterized eukaryotic proteins with large amounts of disorder. We found that virtually all eukaryotic species have proteins with significant amounts of disorder. Several eukaryotes have over 1/3 of their chains with the substantial (>25%) amounts of disorder. Moreover, animals have larger numbers of fully disordered chains relative to

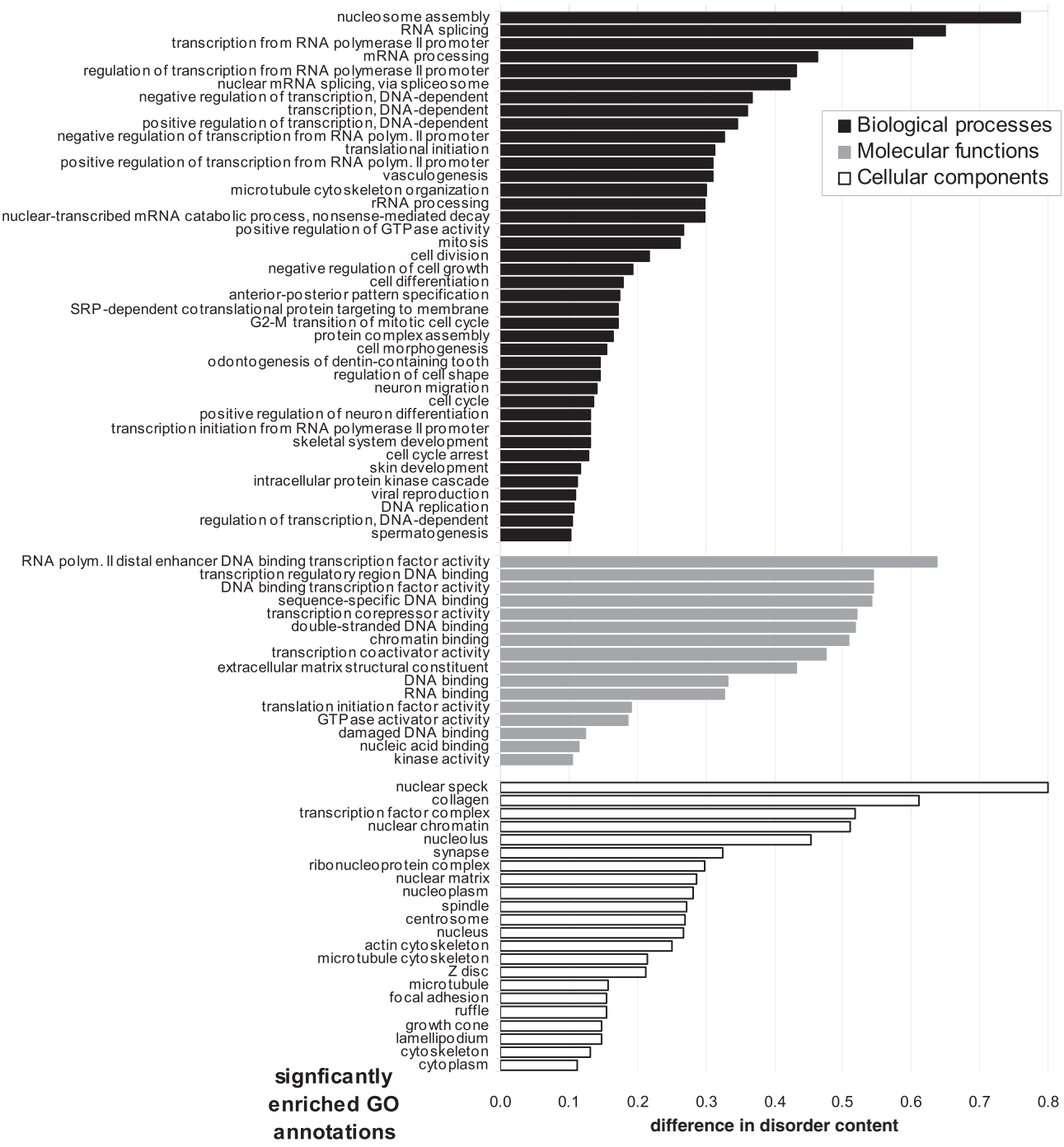

Fig. 8. GO annotations (*y*-axis) of biological processes (black bars), molecular function (gray bars), and cellular components (hollow bars) that are significantly enriched in disorder, *p*-value < 0.05, in *Homo sapiens*. The *x*-axis shows the amount of enrichment in disorder content compared to the chain size-matched content in the human proteome.

other phyla. We show that 45% of the considered species, again with a large fraction of animal species, have 50+ fully disordered chains. Moreover, we demonstrate that large disorder content is primarily found in relatively short (100 or fewer AAs) protein chains, however many of the animal and fungal species also have very large proteins (1000 or more AAs) that are heavily disordered.

Finally, we performed a comprehensive investigation of functional roles of disorder in the human proteome, which doubles the coverage offered by prior studies. We found that 22.5% of the considered GO annotations are enriched in disorder. The disorder is implicated in many of key cellular functions, particularly related to the RNA and DNA interactions, and is preferentially localized in several cellular compartments, with nucleus and ribosome as prime examples.

Acknowledgments

We thank Drs Gang Hu and Kui Wang for providing values of the structural coverage. JY was supported by the University of Alberta Doctoral Recruitment Scholarship and by the Discovery grant to LK. MJM was funded by the University of Alberta Dissertation Scholarship.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.bbapap.2013.05.022.

References

[1] P.E. Wright, H.J. Dyson, Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm, J. Mol. Biol. 293 (1999) 321–331.
[2] V.N. Uversky, The mysterious unfoldome: structureless, underappreciated, yet vital part of any given proteome, J. Biomed. Biotechnol. 2010 (2010) 568068.
[3] A. Dunker, M. Babu, E. Barbar, M. Blackledge, S. Bondos, Z. DosztÃ¡nyi, H. Dyson, J. Forman-Kay, M. Fuxreiter, J. Gsponer, K. Han, D. Jones, S. Longhi, S. Metallo, K. Nishikawa, R. Nussinov, Z. Obradovic, R. Pappu, B. Rost, P. Selenko, V. Subramaniam, J. Sussman, P. Tompa, V. Uversky, What's in a name? Why these proteins are intrinsically disordered, Intrins. Disorder. Proteins 1 (2013) e24157.
[4] J.J. Ward, J.S. Sodhi, L.J. McGuffin, B.F. Buxton, D.T. Jones, Prediction and functional analysis of native disorder in proteins from the three kingdoms of life, J. Mol. Biol. 337 (2004) 635–645.
[5] B. Xue, A.K. Dunker, V.N. Uversky, Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life, J. Biomol. Struct. Dyn. 30 (2012) 137–149.
[6] A.K. Dunker, C.J. Brown, J.D. Lawson, L.M. Iakoucheva, Z. Obradovic, Intrinsic disorder and protein function, Biochemistry 41 (2002) 6573–6582.
[7] L.M. Iakoucheva, C.J. Brown, J.D. Lawson, Z. Obradovic, A.K. Dunker, Intrinsic disorder in cell-signaling and cancer-associated proteins, J. Mol. Biol. 323 (2002) 573–584.
[8] A.K. Dunker, M.S. Cortese, P. Romero, L.M. Iakoucheva, V.N. Uversky, Flexible nets: the roles of intrinsic disorder in protein interaction networks, FEBS J. 272 (2005) 5129–5148.
[9] V.N. Uversky, C.J. Oldfield, A.K. Dunker, Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling, J. Mol. Recognit. 18 (2005) 343–384.
[10] P. Radivojac, L.M. Iakoucheva, C.J. Oldfield, Z. Obradovic, V.N. Uversky, A.K. Dunker, Intrinsic disorder and functional proteomics, Biophys. J. 92 (2007) 1439–1456.
[11] V.N. Uversky, C. Oldfield, U. Midic, H. Xie, B. Xue, S. Vucetic, L.M. Iakoucheva, Z. Obradovic, A.K. Dunker, Unfoldomics of human diseases: linking protein intrinsic disorder with diseases, BMC Genomics 10 (Suppl. 1) (2009) S7.
[12] V.N. Uversky, C.J. Oldfield, A.K. Dunker, Intrinsically disordered proteins in human diseases: introducing the D2 concept, Annu. Rev. Biophys. 37 (2008) 215–246.
[13] B. He, K. Wang, Y. Liu, B. Xue, V.N. Uversky, A.K. Dunker, Predicting intrinsic disorder in proteins: an overview, Cell Res. 19 (2009) 929–949.
[14] Z.-L. Peng, L. Kurgan, Comprehensive comparative assessment of in-silico predictors of disordered regions, Curr. Protein Pept. Sci. 13 (2012) 6–18.
[15] X. Deng, J. Eickholt, J. Cheng, A comprehensive overview of computational protein disorder prediction methods, Mol. Biosyst. 8 (1) (2012) 114–121.
[16] D.T. Jones, J.J. Ward, Prediction of disordered regions in proteins from position specific score matrices, Proteins 53 (Suppl. 6) (2003) 573–578.
[17] Z. Dosztányi, V. Csizmok, P. Tompa, I. Simon, The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins, J. Mol. Biol. 347 (4) (2005) 827–839.
[18] Z.R. Yang, R. Thomson, P. McMeil, R.M. Esnouf, RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins, Bioinformatics 21 (2005) 3369–3376.
[19] A. Schlessinger, G. Yachdav, B. Rost, PROFbval: predict flexible and rigid residues in proteins, Bioinformatics 22 (2006) 891–893.
[20] A. Schlessinger, J. Liu, B. Rost, Natively unstructured loops differ from other loops, PLoS Comput. Biol. 3 (2007) e140.
[21] A. Schlessinger, M. Punta, B. Rost, Natively unstructured regions in proteins identified from contact predictions, Bioinformatics 23 (2007) 2376–2384.
[22] T. Ishida, K. Kinoshita, PrDOS: prediction of disordered protein regions from amino acid sequence, Nucleic Acids Res. 35 (2007) W460–W464.
[23] L.J. McGuffin, Intrinsic disorder prediction from the analysis of multiple protein fold recognition models, Bioinformatics 24 (2008) 1798–1804.
[24] A. Schlessinger, et al., Improved disorder prediction by combination of orthogonal approaches, PLoS One 4 (2009) e4433.
[25] X. Deng, J. Eickholt, J. Cheng, PreDisorder: ab initio sequence-based prediction of protein disordered regions, BMC Bioinforma. 10 (2009) 436.
[26] S. Hirose, K. Shimizu, T. Noguchi, POODLE-I: disordered region prediction by integrating POODLE series and structural information predictors based on a workflow approach, In Silico Biol. 10 (3) (2010) 185–191.
[27] M.J. Mizianty, W. Stach, K. Chen, K.D. Kedarisetti, F.M. Disfani, L. Kurgan, Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources, Bioinformatics 26 (2010) i489–i496.
[28] B. Xue, R.L. Dunbrack, R.W. Williams, A.K. Dunker, V.N. Uversky, PONDR-FIT: a meta-predictor of intrinsically disordered amino acids, Biochim. Biophys. Acta 1804 (4) (2010) 996–1010.
[29] I. Walsh, A.J. Martin, T. Di Domenico, A. Vullo, G. Pollastri, S.C. Tosatto, CSpritz: accurate prediction of protein disorder segments with annotation for homology, secondary structure and linear motifs, Nucleic Acids Res. 39 (2011) W190–W196.
[30] I. Walsh, A.J. Martin, T. Di Domenico, S.C. Tosatto, ESpritz: accurate and fast prediction of protein disorder, Bioinformatics 28 (4) (2012) 503–509.
[31] L.P. Kozlowski, J.M. Bujnicki, MetaDisorder: a meta-server for the prediction of intrinsic disorder in proteins, BMC Bioinforma. 13 (2012) 111.
[32] T. Zhang, E. Faraggi, B. Xue, A.K. Dunker, V.N. Uversky, Y. Zhou, SPINE-D: accurate prediction of short and long disordered regions by a single neural-network based method, J. Biomol. Struct. Dyn. 29 (4) (2012) 799–813.
[33] B. Monastyrskyy, K. Fidelis, J. Moult, A. Tramontano, A. Kryshtafovych, Evaluation of disorder predictions in CASP9, Proteins 79 (2011) 107–118, (Sup10).
[34] M.J. Mizianty, T. Zhang, B. Xue, Y. Zhou, A.K. Dunker, V.N. Uversky, L. Kurgan, In-silico prediction of disorder content using hybrid sequence representation, BMC Bioinforma. 12 (2011) 245.
[35] T. Le Gall, P.R. Romero, M.S. Cortese, V.N. Uversky, A.K. Dunker, Intrinsic disorder in the Protein Data Bank, J. Biomol. Struct. Dyn. 24 (4) (2007) 325–342.
[36] Z. Dosztányi, J. Chen, A.K. Dunker, I. Simon, P. Tompa, Disorder and sequence repeats in hub proteins and their implications for network evolution, J. Proteome Res. 5 (11) (2006) 2985–2995.
[37] H. Hegyi, L. Buday, P. Tompa, Intrinsic structural disorder confers cellular viability on oncogenic fusion proteins, PLoS Comput. Biol. 5 (10) (2009) e1000552.
[38] Z.-L. Peng, M.J. Mizianty, B. Xue, L. Kurgan, V.N. Uversky, More than just tails: intrinsic disorder in histone proteins, Mol. Biosyst. 8 (2012) 1886–1901.
[39] B. Xue, R.W. Williams, C.J. Oldfield, A.K. Dunker, V.N. Uversky, Archaic chaos: intrinsically disordered proteins in Archaea, BMC Syst. Biol. 4 (Suppl. 1) (2010) S1.
[40] H. Xie, S. Vucetic, L.M. Iakoucheva, C.J. Oldfield, A.K. Dunker, Z. Obradovic, V.N. Uversky, Functional anthology of intrinsic disorder. 3. Ligands, post-translational modifications, and diseases associated with intrinsically disordered proteins, J. Proteome Res. 6 (2007) 1917–1932.
[41] S. Vucetic, H. Xie, L.M. Iakoucheva, C.J. Oldfield, A.K. Dunker, Z. Obradovic, V.N. Uversky, Functional anthology of intrinsic disorder. 2. Cellular components, domains, technical terms, developmental processes, and coding sequence diversities correlated with long disordered regions, J. Proteome Res. 6 (2007) 1899–1916.
[42] A. Lobley, M.B. Swindells, C.A. Orengo, D.T. Jones, Inferring function using patterns of native disorder in proteins, PLoS Comput. Biol. 3 (8) (2007) e162.
[43] U. Midic, C. Oldfield, A.K. Dunker, Z. Obradovic, V.N. Uversky, Protein disorder in the human diseasome: unfoldomics of human genetic diseases, BMC Genomics 10 (Suppl. 1) (2009) S12.
[44] R. Linding, L.J. Jensen, F. Diella, P. Bork, T.J. Gibson, R.B. Russell, Protein disorder prediction: implications for structural proteomics, Structure 11 (11) (2003) 1453–1459.
[45] W.N. Price, Y. Chen, S.K. Handelman, H. Neely, P. Manor, R. Karlin, R. Nair, J. Liu, M. Baran, J. Everett, S.N. Tong, F. Forouhar, S.S. Swaminathan, T. Acton, R. Xiao, J.R. Luft, A. Lauricella, G.T. DeTitta, B. Rost, G.T. Montelione, J.F. Hunt, Understanding the physical properties that control protein crystallization by analysis of large-scale experimental data, Nat. Biotechnol. 27 (2009) 51–57.
[46] M.J. Mizianty, L. Kurgan, Sequence-based prediction of protein crystallization, purification, and production propensity, Bioinformatics 27 (13) (2011) i24–i33.
[47] C.J. Oldfield, B. Xue, Y.Y. Van, E.L. Ulrich, J.L. Markley, A.K. Dunker, V.N. Uversky, Utilization of protein intrinsic disorder knowledge in structural proteomics, Biochim. Biophys. Acta 1834 (2) (2013) 487–498.
[48] F. Miri Disfani, W.L. Hsu, M.J. Mizianty, C.J. Oldfield, B. Xue, A.K. Dunker, V.N. Uversky, L. Kurgan, MoRFpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins, Bioinformatics 28 (12) (2012) i75–i83.
[49] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The Protein Data Bank, Nucleic Acids Res. 28 (2000) 235–242.
[50] M. Sickmeier, J.A. Hamilton, T. LeGall, V. Vacic, M.S. Cortese, A. Tantos, B. Szabo, P. Tompa, J. Chen, V.N. Uversky, Z. Obradovic, A.K. Dunker, DisProt: the database of disordered proteins, Nucleic Acids Res. 35 (2007) D786–D793.
[51] F.L. Sirota, H.S. Ooi, T. Gattermayer, G. Schneider, F. Eisenhaber, S. Maurer-Stroh, Parameterization of disorder predictors for large-scale applications requiring high specificity by using an extended benchmark dataset, BMC Genomics 11 (Suppl. 1) (2010) S15.
[52] J.C. Wootton, S. Federhen, Statistics of local complexity in amino acid sequences and sequence databases, Comput. Chem. 17 (1993) 149–163.
[53] S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama, M. Kanehisa, AAindex: amino acid index database, progress report 2008, Nucleic Acids Res. 36 (2008) D202–D205.
[54] P. Radivojac, Z. Obradovic, D.K. Smith, G. Zhu, S. Vucetic, C.J. Brown, J.D. Lawson, A.K. Dunker, Protein flexibility and intrinsic disorder, Protein Sci. 13 (1) (2004) 71–80.
[55] V.N. Uversky, J.R. Gillespie, A.L. Fink, Why are "natively unfolded" proteins unstructured under physiologic conditions? Proteins 41 (3) (2000) 415–427.
[56] V.N. Uversky, What does it mean to be natively unfolded? Eur. J. Biochem. 269 (1) (2002) 2–12.
[57] A.K. Dunker, J.D. Lawson, C.J. Brown, R.M. Williams, P. Romero, J.S. Oh, C.J. Oldfield, A.M. Campen, C.M. Ratliff, K.W. Hipps, J. Ausio, M.S. Nissen, R. Reeves, C. Kang, C.R. Kissinger, R.W. Bailey, M.D. Griswold, W. Chiu, E.C. Garner, Z. Obradovic, Intrinsically disordered protein, J. Mol. Graph. Model. 19 (1) (2001) 26–59.
[58] A. Mohan, C.J. Oldfield, P. Radivojac, V. Vacic, M.S. Cortese, A.K. Dunker, V.N. Uversky, Analysis of molecular recognition features (MoRFs), J. Mol. Biol. 362 (2006) 1043–1059.
[59] P. Romero, Z. Obradovic, X. Li, E.C. Garner, C.J. Brown, A.K. Dunker, Sequence complexity of disordered protein, Proteins 42 (1) (2001) 38–48.
[60] UniProt Consortium, Reorganizing the protein space at the Universal Protein Resource (UniProt), Nucleic Acids Res. 40 (2012) D71–D75.
[61] D. Vitkup, E. Melamud, J. Moult, C. Sander, Completeness in structural genomics, Nat. Struct. Biol. 8 (2001) 559–566.
[62] B. Mészáros, I. Simon, Z. Dosztányi, Prediction of protein binding regions in disordered proteins, PLoS Comput. Biol. 5 (5) (2009) e1000376.
[63] E. Schad, P. Tompa, H. Hegyi, The relationship between proteome size, structural disorder and organism complexity, Genome Biol. 12 (12) (2011) R120.
[64] R. Pancsa, P. Tompa, Structural disorder in eukaryotes, PLoS One 7 (4) (2012) e34687.
[65] M. Howell, R. Green, A. Killeen, L. Wedderburn, V. Picascio, A. Rabionet, Z. Peng, M. Larina, B. Xue, L. Kurgan, V.N. Uversky, Not that rigid midgets and not so flexible giants: on abundance and roles of intrinsic disorder in short and long proteins, J. Biol. Syst. 20 (2012) 471–511.
[66] Gene Ontology Consortium, Gene ontology: tool for the unification of biology, Nat. Genet. 25 (1) (2000) 25–29.