PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Genome-scale prediction of proteins with long intrinsically disordered regions

Zhenling Peng, Marcin J. Mizianty, and Lukasz Kurgan*

Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Canada

ABSTRACT

Proteins with long disordered regions (LDRs), defined as having 30 or more consecutive disordered residues, are abundant in eukaryotes, and these regions are recognized as a distinct class of biologically functional domains. LDRs facilitate various cellular functions and are important for target selection in structural genomics. Motivated by the lack of methods that directly predict proteins with LDRs, we designed Super-fast predictor of proteins with Long Intrinsically DisordERed regions (SLIDER). SLIDER utilizes logistic regression that takes an empirically chosen set of numerical features, which consider selected physicochemical properties of amino acids, sequence complexity, and amino acid composition, as its inputs. Empirical tests show that SLIDER offers competitive predictive performance combined with low computational cost. It outperforms, by at least a modest margin, a comprehensive set of modern disorder predictors (that can indirectly predict LDRs) and is 16 times faster compared to the best currently available disorder predictor. Utilizing our time-efficient predictor, we characterized abundance and functional roles of proteins with LDRs over 110 eukaryotic proteomes. Similar to related studies, we found that eukaryotes have many (on average 30.3%) proteins with LDRs with majority of proteomes having between 25 and 40%, where higher abundance is characteristic to proteomes that have larger proteins. Our first-of-its-kind large-scale functional analysis shows that these proteins are enriched in a number of cellular functions and processes including certain binding events, regulation of catalytic activities, cellular component organization, biogenesis, biological regulation, and some metabolic and developmental processes. A webserver that implements SLIDER is available at http://biomine.ece.ualberta.ca/SLIDER/.

Proteins 2014; 82:145–158.

Key words: intrinsic disorder; long disordered regions; disorder prediction; high-throughput prediction; eukaryotes.

INTRODUCTION

Intrinsically disordered proteins (IDPs) and intrinsically disordered regions (IDRs) in protein chains are characterized by lack of stable tertiary structure under physiological conditions *in vitro*.[1] They are found across all domains of life, with archaean and bacterial species having relatively low amounts of disorder and eukaryotes that are strongly enriched in disorder.[2–4] IDPs and IDRs implement important cellular functions,[5–10] and their prevalence was implicated in various human diseases.[11–13] However, currently only a limited number of disordered proteins were characterized experimentally,[14] and these efforts lag behind the rapidly accumulating number of protein chains. This motivates development of computational methods that perform accurate and high-throughput prediction of disorder.

A number of studies have shown that IDPs and IDRs have unique sequence signatures, i.e., disorder is frequently observed in regions with low complexity, low content of hydrophobic amino acids, high content of polar and net-charged residues, in regions that lack secondary structures and that have unique evolutionary and solvent accessibility profiles.[7,15–17] This suggests that disorder is predictable from the protein sequence and enables development of computational approaches for the sequence-derived prediction of disorder. Many such predictors have been developed in the past two decades.[18–20] Since 2002, they are being continually assessed in the biannual CASP (Critical Assessment of Techniques for protein Structure Prediction) experiments.[21–25]

Grant sponsor: Alberta Innovates Graduate Student Scholarship; University of Alberta; Natural Sciences and Engineering Research Council of Canada.
*Correspondence to: Lukasz Kurgan, 9107, 116 Street (ECERF building), Edmonton, Alberta, Canada T6G 2V4. E-mail: lkurgan@ece.ualberta.ca
Received 2 May 2013; Accepted 6 June 2013
Published online 24 June 2013 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/prot.24348

Although these methods predict the disorder relatively well at the residue level,[19,25] they are less accurate at the sequence level. Specifically, they were shown to over- or under-estimate the overall amount of disorder in a given chain[19,26] and to provide a relatively low predictive performance for the prediction of long disordered segments.[19,25] Furthermore, most of these methods perform multiple sequence alignment with PSI-BLAST,[27] which is relatively time consuming. A modern desktop computer takes about 5 minutes to run the alignment for an average-sized protein chain with about 300 residues. This translates to a prohibitive estimate of 70,000*5 = 350,000 minutes, which equals 243 days to calculate these alignments for a human proteome. A few fast predictors, such as ESpritz,[28] IUPred,[29,30] and VSL2B,[31] are available but their predictive quality and computational costs require further improvements, which we demonstrate in our empirical analysis.

In this work, we focus on the prediction of proteins with long disordered regions (LDRs), which are defined as having 30 or more consecutive residues in length.[3,4,32–34] By conservative estimates, about 10–35% of prokaryotic and about 15–45% of eukaryotic proteins contain these long disordered regions,[3,32] with over 40% in the human proteome.[35] The number of proteins with LDRs was estimated to be an order of magnitude higher in eukaryotes than in archaea and bacteria.[36] They were found in various protein families, with examples being transmembrane proteins[37] and spliceosome proteins.[38] Moreover, LDRs are recognized as a distinct class of biologically functional protein domains, which points to their important role as functional elements.[35,39] Importantly, knowledge of LDRs finds practical applications, as they are implicated in protein–protein recognition and are important for target selection in structural genomics.[39–42] Motivated by the importance and abundance of LDRs, lack of methods that directly predict proteins with LDRs, and the fact that current disorder predictors (which can be indirectly used to find LDRs) are deficient in speed and accuracy, our aim was to develop an accurate and fast predictor of proteins with LDRs. Our Super-fast predictor of proteins with Long Intrinsically DisordERed regions (SLIDER) is characterized by several advantages: (1) low computational cost (our method is at least an order of magnitude faster than the existing methods; it predicts an entire human proteome in about 30 minutes on a desktop computer); (2) comprehensive design (we utilize feature selection to design a well-performing set of descriptors that are generated by combining composition and physiochemical properties of amino acids and sequence complexity, and a fast, empirically designed logistic regression-based prediction model); and (3) good predictive performance (SLIDER offers at least modest improvements when compared with a comprehensive set of modern disorder predictors on a large benchmark set that shares low similarity to the SLIDER's training dataset). We also applied our predictor to analyze abundance and to perform first-of-its-kind large-scale functional characterization of chains with LDRs in 110 eukaryotic proteomes. Our results on the abundance are in agreement with previous studies, while the functional analysis reveals that LDRs are primarily involved in transcription, enzyme regulation, and various binding events.

MATERIALS AND METHODS

Datasets

We utilized a recently proposed MxD dataset[43] to design and test our predictor. The disorder in this dataset is defined based on curated annotations collected from release 4.9 of the DisProt database[14] and annotations based on REMARK 465 utilizing structures from the Protein Data Bank (PDB),[44] which is consistent with protocols used in CASP. Furthermore, the annotations for chains from DisProt were enriched using the SL dataset-based procedure,[45] similar to that in several related recent studies.[26,46–49] Each protein was classified as either having or not having the long disordered region; residues in chains from DisProt that lack annotations were ignored when classifying the chains. The original set of 514 chains was reduced to 494 proteins since we had to remove several chains that could not be predicted by the evaluated disorder predictors. These 494 chains were randomly partitioned into equal-sized TRAINING and TEST datasets, each with 247 proteins including 130 and 128 chains with LDRs, respectively. Since the chains in the MxD set are characterized by pairwise sequence identity below 25%,[43] the TRAINING and TEST datasets also share this low level of similarity. We could not use the benchmark sets from the recent CASP9 and CASP10 experiments since they include only 8 and 10 LDRs, respectively, which would not allow for a statistically sound evaluation. The TRAINING and TEST datasets are available at http://biomine.ece.ualberta.ca/SLIDER/

We extracted all 110 fully sequenced eukaryotic proteomes (1,901,810 proteins) from the release 2011_08 of UniProt.[50] These proteomes are predicted with SLIDER to investigate abundance and functional roles of putative proteins with LDR in eukaryotes. We also used this dataset to perform large-scale assessment of the runtime of SLIDER.

Evaluation criteria and statistical significance

The prediction consists of a binary label (protein with or with no LDR) together with a numeric score that quantifies propensity of the input chain to have LDRs.

The binary predictions were assessed using four measures:

$$\text{Accuracy} = (TP + TN)/(TP + FP + TN + FN)$$

$$\text{Mathews Correlation Coefficient (MCC)} = (TP * TN - FP * FN)/\text{sqrt}((TP + FP)(TP + FN)(TN + FP)(TN + FN))$$

$$\text{Sensitivity} = TP/(TP + FN)$$

$$\text{Specificity} = TN/(TN + FP)$$

where TP (true positive) is the count of correctly predicted proteins with LDRs, TN (true negative) is the number of correctly predicted proteins without LDRs, FP (false positive) is the number of chains that do not have LDRs but were predicted to have them, and FN (false negative) is the count of proteins that have LDRs but were predicted not to have them. MCC values range between −1 and 1, and they are equal to zero when all proteins in a given dataset are predicted to have the same outcome.

The predicted propensities were evaluated using the receiver operating characteristic (ROC) curves. For a propensity threshold P that ranges between 0 and 1, the proteins with predicted propensity equal or greater than P are assumed to be positive (having LDRs), and all other chains are set as negative (having no LDRs). Next, the TP-rate = TP/(TP+FN) and the FP-rate = FP/(FP+TN) are used to draw the ROC curve by varying the values of P. We used the area under the ROC curve (AUC) to quantify the predictive quality, where higher values indicate better predictions.

We also assessed statistical significance of improvements, which are measured with MCC and AUC, offered by our method when compared to the other predictors. First, we randomly selected 10 sets of 100 protein chains from the TEST dataset and computed the MCC and AUC of each considered predictor on each of the 10 protein sets. Next, we evaluated significance of the differences in MCC/AUC between these 10 paired results, i.e., SLIDER's results against results of another predictor. If these MCC/AUC values follow normal distribution, as tested using Shapiro-Wilk test[51] at the 0.05 significance, then we utilized paired t-test; otherwise we used the Wilcoxon rank sum test.[52] The MCC/AUC of a given predictor is assumed to be equivalent to that of SLIDER if the resulting P-value >0.05; otherwise we assumed that the difference is significant.

Design of the predictive model

The design of the model was performed in four steps utilizing the TRAINING dataset. First, we investigated relation between various physiochemical properties of amino acids (AAs) and the native annotations of the LDRs. This led to the empirical selection of certain properties that are useful for the prediction of proteins with LDRs. Second, we developed a numerical (feature-based) representation of the input protein chains using the composition of the sequence and these selected properties. Third, we performed correlation-based feature selection to remove irrelevant (to the prediction of chains with LDRs) and redundant (with each other) features. Fourth, we empirically selected and parameterized prediction model that uses the selected features.

Selection of relevant physiochemical properties of amino acids

Several physicochemical properties of AAs, such as flexibility, solvent accessibility, net charge, hydrophobicity, etc., were successfully used to implement existing disorder predictors.[7,17,53,54] This motivated our approach, in which we investigated whether AA indices from the AAindex database are useful to predict chains with LDRs. We collected all 544 indices from version 9.1 of AAindex[55,56] to comprehensively cover physicochemical properties of AAs; 13 indices were removed since they include unknown/missing values.

Since some of these indices are likely irrelevant to the prediction of LDRs, we filtered them by empirically assessing their relevance using the TRAINING dataset. For a given index, we calculated the average of its values for AAs in LDRs and average of its values in the remaining residues for the chains from the TRAINING dataset. We evaluated whether these averages are significantly different by repeating these calculations for 10 sets of randomly selected 100 proteins from the TRAINING dataset. If these averages follow normal distribution, which was tested with the Shapiro-Wilk test at the 0.05 significance, then we used paired t-test to evaluate significance; otherwise we used the Wilcoxon rank sum test. We removed all indices for which the resulting P-value >0.05. Consequently, we kept 451 indexes which are characterized by significant differences between AAs in LDRs and other AAs. Next, we discarded redundant indices, i.e., those that are similar to each other. We computed the Pearson Correlation Coefficient (PCC) between values of each pair of the remaining indices to estimate the redundancy. We grouped them together such that each pair of indices in a given group has PCC > 0.7, and for each group we retain only one index that has the smallest P-value. As a result, we selected 48 AA indices.

Sequence representation

Besides the physicochemical properties, previous studies have shown bias of particular AAs toward disordered conformations and demonstrated that complexity of the underlying sequence is related with propensity for disorder.[1,57–59] Importantly, these sequence-derived

characteristics can be computed very quickly, which facilitates our goal to keep the computational costs low. The above motivates the use of the selected AA indices, sequence complexity, and the AA composition to develop feature-based/numerical representation of the input protein sequence. We considered the following four sets of features:

1. AA composition, which is defined as the ratio of the number residues of a given AA type among all residues in the input protein chain (20 features).
2. Features based on the annotation of the low/high complexity regions generated by the SEG algorithm[60,61] (10 features). We calculated the number of AAs in the low/high complexity regions (two features), the number of low/high complexity segments with at least four consecutive residues (two features), and the average and maximum length of the low/high complexity segments (2*2 = 4 features); these features were normalized by the protein length. Furthermore, we used a sliding window with size of 30 (which corresponds to the minimal length of the LDRs) to count how many of the resulting 30-residues-long segments are composed entirely from residues assigned as either low or high complexity; these counts were normalized by the total number 30-residues-long segments in a given sequence (two features).
3. Features based on the selected AA indices/physicochemical properties (144 features). We computed three features for each of the 48 selected AA indices. First, we calculated the average value of a given AA index over all residues in the input protein chain. The other two features correspond to the minimum and maximum averages among the values calculated using a sliding window with size of 30. The latter quantifies particular bias in a given chain to have 30-residues-long segments with extreme values of the selected physicochemical properties.
4. Hybrid features that combine AA compositions, the selected AA indices, and the annotation of complexity regions (328 features). We computed AA composition for residues in the high and low complexity regions, respectively, that were generated by the SEG method (20*2 = 40 features). We also calculated average values of each of the 48 selected AA indices using features defined in the feature set 3 in the high and low complexity regions, respectively (144*2 = 288 features).

All together we generated $20 + 10 + 144 + 328 = 502$ features.

Feature selection

Some of the considered 502 features could be irrelevant to the prediction of proteins with LDRs and/or could be redundant with each other. Thus, for each feature we investigated its relevance to our prediction task by computing average point-biserial correlation[62] between values of this feature and the native, binary labels of proteins with/without LDRs. The average is based on five point-biserial correlations that correspond to five training folds that are generated utilizing five-fold cross-validation on the TRAINING dataset. We removed features characterized by low, below 0.2, correlations. Next, we computed PCC values for all pairs of the remaining features to remove redundancy. We used six thresholds to define redundant features, i.e., a given feature is assumed to be redundant if it has PCC with another feature that is greater than 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9, respectively. We grouped the features together such that each pair of features in a given group has PCC greater than a given threshold and we selected one feature, the one with the highest point-biserial correlation, from each group. As a result, we obtained six sets of features, depending on the value of the redundancy threshold. Figure 1 shows (see the gray bars) that the sizes of the six feature sets decrease proportionally with the value of the threshold, from 124 features for the threshold = 0.9 to 19 features for the threshold = 0.4.

Selection and parameterization of prediction model

We considered two popular classification algorithms, Support Vector Machine (SVM) and ridge logistic regression, to generate our prediction model. Moreover, we tested two commonly used kernel functions, Radial Basis Function (RBF) and polynomial, to implement SVM. Each of the three classifiers, including two versions of SVM and the ridge regression, was parameterized. The parameterization was performed using grid search in which we vary the values of the complexity constant $C = 2^x$, where $x = (-8, -7, \ldots, 8)$ and the degree of polynomial $d = (1, 2)$ for the SVM with the polynomial kernel; complexity constant $C = 2^x$ and $\gamma = 2^x$ of the RBF function for the SVM with the FBF kernel, and the ridge parameter $r = 10^z$, where $z = (-11, -9, \ldots, 4)$, for the ridge logistic regression. The parameters that correspond to the highest value of AUC based on the five-fold cross-validation on the TRAINING dataset were selected. This parameterization was performed for each of the six features sets generated based on different values of the redundancy threshold for a total of 18 setups; the results are summarized in Figure 1. The ridge logistic regression provides slightly higher values of the cross-validated AUC when compared with both versions of SVM, except when using low values of the redundancy threshold. Moreover, the regression is also slightly faster to compute compared to the SVM model that utilizes the same feature set. Overall, the highest AUC of 0.853 was obtained with the ridge logistic regression and the 51 features selected using the threshold of 0.6. This setup constitutes the design of our SLIDER predictor. The ridge logistic regression generates numeric propensities of a given

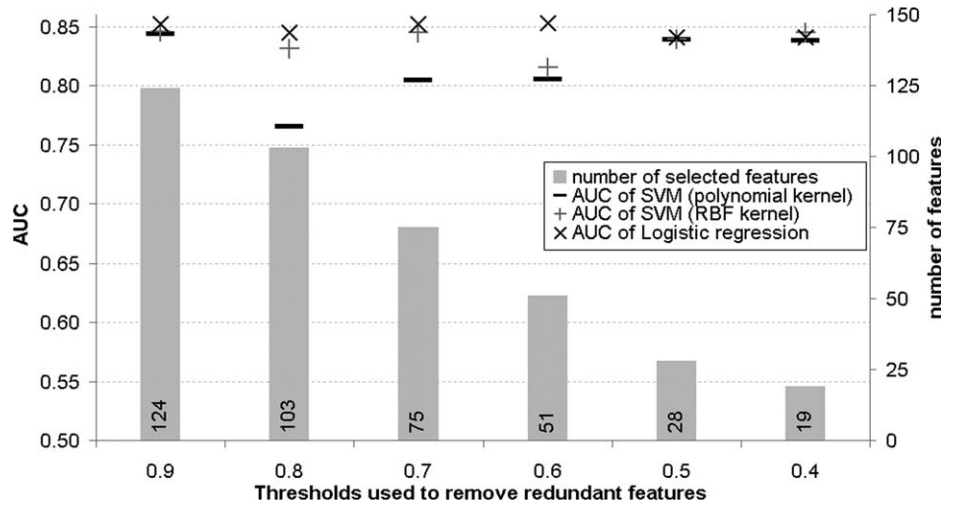

Figure 1

Summary of results obtained during feature selection and selection and parameterization of the prediction model. The *x*-axis denotes the threshold used to remove redundant features. The gray bars denote the number of selected features, which are quantified using the *y*-axis on the right and shown using numbers at the bottom of the bars. The markers denote the predictive quality, measured using AUC shown on the *y*-axis on the right, for a given prediction model. All results are based on the five-fold cross-validation on the TRAINING dataset.

input chain to have LDR. The binary predictions were defined based on thresholding these propensities. A given chain is predicted to have LDR if the propensity >0.538; otherwise it is predicted not to have LDR. This threshold corresponds to the maximal value of MCC based on the predictions from the five-fold cross-validation on the TRAINING dataset. The selected 51 features utilize 30 different AA indices that quantify a variety of physicochemical properties of AAs. These properties include secondary structures, solvent accessibility, hydrophobicity, and flexibility that have been utilized in previous related works; and some potentially new, in the context of the disorder prediction, factors related to the propensity to form transmembrane regions, heat capacity, unfolding energy, and electron-ion interaction potentials. This shows that our predictor applies novel sequence-derived markers to find proteins with LDRs. The SLIDER method is available, as an easy to use webserver, at http://biomine.ece.ualberta.ca/SLIDER/.

RESULTS AND DISCUSSION

Comparative evaluation of predictive quality

The predictive quality of SLIDER is compared with a comprehensive set of 23 modern disorder predictors on the TEST dataset (see Table I). These predictors include 13 methods that were evaluated in a recent review[19] and other recently published predictors, such as PreDisorder,[63] PrDos,[64] two version of CSpritz: short and long,[46] and six versions ESpritz: 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).[28] They include publicly available versions of the top three disorder predictors according to AUC and MCC measures from the CASP9 experiment: PrDOS, DISOPRED, PreDisorder (also called MULTICOM), and MFDp.[25] We divide these methods into those that are based on a single sequence, which are fast to compute, and those that utilize multiple sequence alignment, which comes with a higher computational cost. We convert the residue level predictions from the considered 23 methods into the per-sequence predictions of proteins with LDRs as follows. A given chain is predicted to have LDR if the per-residue prediction includes at least one LDR, i.e., segment of 30 or more consecutive predicted disordered residues; otherwise it is predicted not to have LDR. For chains with LDRs, the numeric score that quantifies propensity to have LDRs was calculated as maximum among the average per-residues probabilities generated by a given prediction for 30 AA long sliding windows, i.e., highest average propensity for disorder among all LDRs. For chains without LDR, the propensity is defined as minimum among the average per-residues probabilities using the same sliding windows. This approach maximizes the predictive quality of the per-residue predictors.

Table I shows that SLIDER obtains the highest values of MCC and AUC. The magnitude of the improvements over the other methods ranges from modest to relatively large. SLIDER improves MCC by 100%*(0.63–0.61)/0.61 = 3.3% and AUC by 100%*(0.87–0.86)/0.86 = 1.2%

Table I
Comparison of Predictive Quality of SLIDER and 23 Modern Disorder Predictors on the TEST Dataset

Predictors		Reference	Accuracy	Sensitivity	Specificity	MCC	Sig.	AUC	Sig.
Single-sequence predictors	SLIDER	This paper	**0.81**	0.73	0.90	**0.63**		**0.87**	
	VSL2B	31	**0.81**	**0.86**	0.75	0.61	=	0.86	=
	ESpritz Cx-S_w	28	0.76	0.76	0.76	0.52	+	0.79	++
	IUPred short	29	0.76	0.63	0.91	0.55	=	0.77	++
	Espritz NMR-S_w	28	0.70	0.73	0.67	0.40	++	0.77	++
	IUPred long	29	0.74	0.57	0.93	0.54	+	0.76	++
	Espritz Cx-FPR	28	0.74	0.63	0.86	0.49	+	0.76	++
	Espritz NMR-FPR	28	0.72	0.63	0.82	0.45	+	0.75	++
	Espritz DP-S_w	28	0.66	0.82	0.49	0.33	++	0.75	++
	Espritz DP-FPR	28	0.71	0.52	0.92	0.47	++	0.67	++
Predictors that utilize multiple sequence alignment	Cspritz long	46	**0.81**	0.83	0.79	0.62	=	0.84	+
	Cspritz short	46	0.78	0.80	0.75	0.55	+	0.83	+
	RONN	65	0.78	0.80	0.76	0.55	=	0.82	+
	DISOCLUST	66	0.75	0.80	0.70	0.51	+	0.82	++
	PreDisorder	63	0.78	0.76	0.81	0.56	=	0.82	+
	MFDp	43	0.78	0.80	0.76	0.56	=	0.81	++
	PONDR-FIT	67	0.78	0.68	0.89	0.58	=	0.81	++
	DISOPRED2	68	0.78	0.68	0.88	0.57	=	0.80	++
	PrDos	64	0.74	0.55	0.94	0.53	+	0.78	++
	NORSnet	69	0.72	0.49	**0.97**	0.53	+	0.77	++
	Profbval	70	0.70	0.73	0.66	0.40	++	0.77	++
	MD	71	0.73	0.66	0.82	0.48	++	0.72	++
	DISpro	72	0.68	0.40	0.97	0.45	++	0.69	++
	Ucon	73	0.70	0.45	0.97	0.48	+	0.69	++

The highest values for each quality index are shown in bold font. The "Sig." columns give results of the test of significance of the differences in MCC and AUC between SLIDER and a given predictor. The test compares results obtained with 10 randomly selected sets of 100 chains from the TEST dataset; "+" and "++" indicate that improvements offered by SLIDER were significant with *P*-value <0.05 and <0.001, respectively; "=" denotes that the differences were not significant.

compared to the second-best in AUC VLS2B, and by 1.6% in MCC and by 3.6% in AUC compared with the second-best in MCC CSpritz long. These improvements were found to be statistically significant for all other predictors except VLS2B in case of AUC, and for 15 out of 23 methods when using MCC. Our predictor is characterized by relatively high specificity coupled with strong levels of sensitivity. This means that it relatively rarely generates false positives (i.e., relatively few incorrect predictions among the chains predicted to have LDRs) and that it correctly predicts majority (73%) of the native chains with LDRs. Figure 2 gives the ROC curves for the four top-performing according to AUC methods: SLIDER, VSL2B, and two versions of CSpritz. SLIDER provides high TP-rates for low FP-rates (i.e., can find a high-quality subset of native chains with LDRs), which are comparable with the results from VSL2B. It can also find the highest number of native chains with LDRs for higher values of the FP-rate, above 0.45.

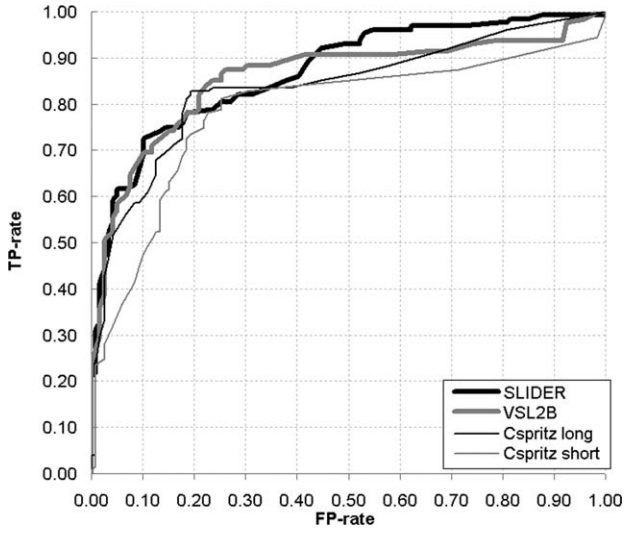

Figure 2
ROC curves for the four predictors with the highest AUC values: SLIDER, VSL2B, CSpritz long, and CSpritz short on the TEST dataset.

This comparative analysis revealed that our specialized predictor of chains with LDR offers competitive predictive performance compared to a comprehensive set of state-of-the-art generic, per-residue predictors. Moreover, our predictions are very fast to compute, which gives a substantial advantage when considering large-scale applications, which we demonstrate next.

Comparative evaluation of runtime

Our empirical analysis demonstrates that several methods obtain comparable to SLIDER levels of predictive quality, i.e., high AUC >0.8 combined with high MCC >0.55; they include VSL2B,[31] CSpritz,[46] PONDR-FIT,[67] PreDisorder[63], and MFDp.[43] We compared runtime of SLIDER with these top-performing predictors. CSpritz,

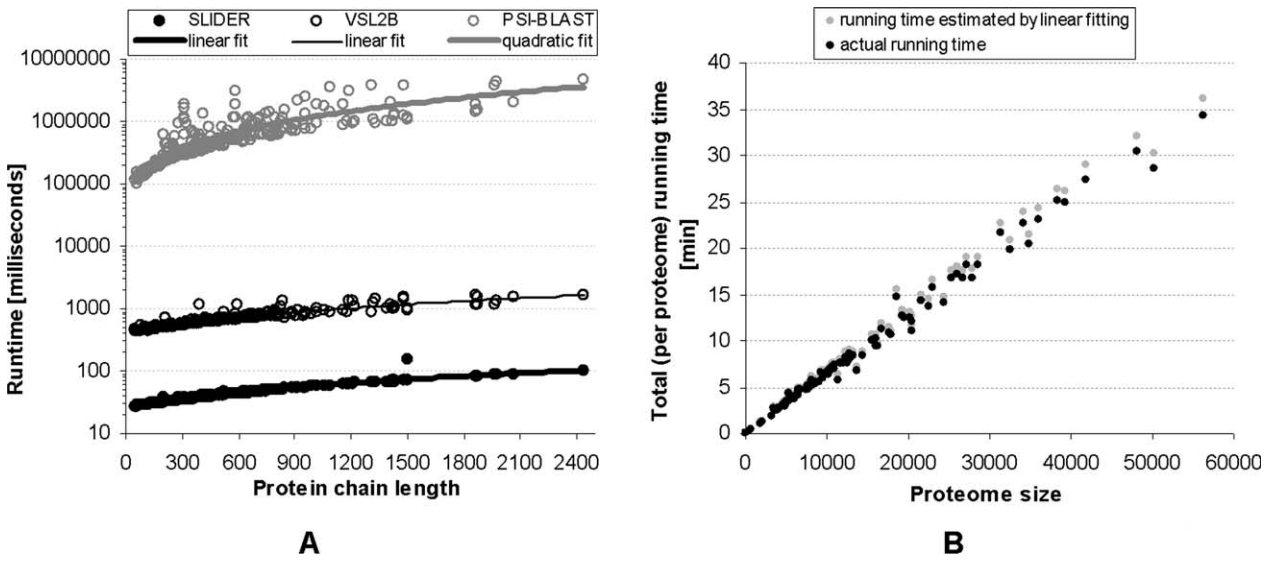

Figure 3

Comparison of runtime. Panel **A** shows relationships between the length of proteins chains (*x*-axis) and the runtime in milliseconds (*y*-axis in logarithmic scale) computed for individual chains from the TEST dataset using a modern desktop computer for SLIDER (solid black markers), VSL2B (hollow black markers), and one iteration (j=1) of PSI_BLAST (hollow gray markers). Linear fits for these relations are shown for SLIDER (thick black line) and VSL2B (thin black line); a quadratic fit is shown for the PSI_BLAST (thick gray line). Panel **B** gives relationship between the proteome size and the total actual (black markers) and estimated based on linear fit from Panel A (gray markers) runtime in minutes to complete predictions with SLIDER on a modern desktop computer for each of the considered 110 eukaryotic proteomes.

PONDR-FIT, PreDisorder, and MFDp utilize PSI-BLAST[27] and thus we estimated their runtime by the time to run PSI-BLAST with one iteration (i.e., $j = 1$) against the nr database. The runtimes of SLIDER, VSL2B, and one-iteration of PSI-BLAST that were calculated on the TEST dataset using a modern desktop computer are compared in Figure 3(A). We focus our analysis on relative differences in the runtime, rather than on the absolute values, since these are hardware independent. On average, over all considered chains, SLIDER is 16.2 times faster than VSL2B and over three orders of magnitude faster than PSI-BLAST; these differences are consistent across all chain sizes. Prediction for a single chain with SLIDER takes between 25 and 100 milliseconds, depending on the chain length. The runtimes of SLIDER and VSL2B grow linearly with the chain size, in contrast to PSI_BLAST that registers a quadratic increase. Comparison of the increase between the runtime for shortest (<100 AAs) and longest (>1000 AAs) chains for SLIDER and VLS2B shows a modest 2.6-fold increase, compared to the 15.1-fold increase when using PSI-BLAST.

We further investigate these results by estimating the runtime to perform predictions on 110 complete eukaryotic proteomes, i.e., 1,901,810 proteins. We used linear fitting from Figure 3(A) to estimate the runtime for SLIDER and VLS2B and the quadratic fitting for the PSI-BLAST. Both, linear and quadratic fits, provide good approximations of the measured data, see Figure 3(A). Figure 3(B), which compares the actual and approximated (using the linear fit) total runtime of SLIDER for each of the 110 proteomes, confirms that the linear fit provides accurate runtime estimates; the average absolute error in the estimate is 26 seconds compared to an average per-proteomes runtime of 8 minutes and 46 seconds. The total runtimes for the considered 1.9 million chains for SLIDER, VSL2B, and one-iteration PRI-BLAST are approximately 21 hours, 14 days, and 36 years, respectively. We note that the considered 110 eukaryotic proteomes constitute less than 0.02% of the total number of known (to date) eukaryotic species, which exceeded 554,000 in a recent version of UniProt. To sum up, our results demonstrate that SLIDER provides a fast and accurate prediction of chains with LDRs that is required to perform analysis in the high-throughput, proteomic-scale setting.

Abundance of chains with LDRs in eukaryotes

Using SLIDER, we assessed and characterized abundance of proteins with LDRs in 110 fully sequenced eukaryotic proteomes collected from UniProt[50] and contrasted our results with similar recent analyses. An early work investigated abundance of proteins with LDRs in a limited set of five eukaryotic species.[3] This was followed by more recent contributions that include analysis of 67 eukaryotes[4] and the largest to date investigation that

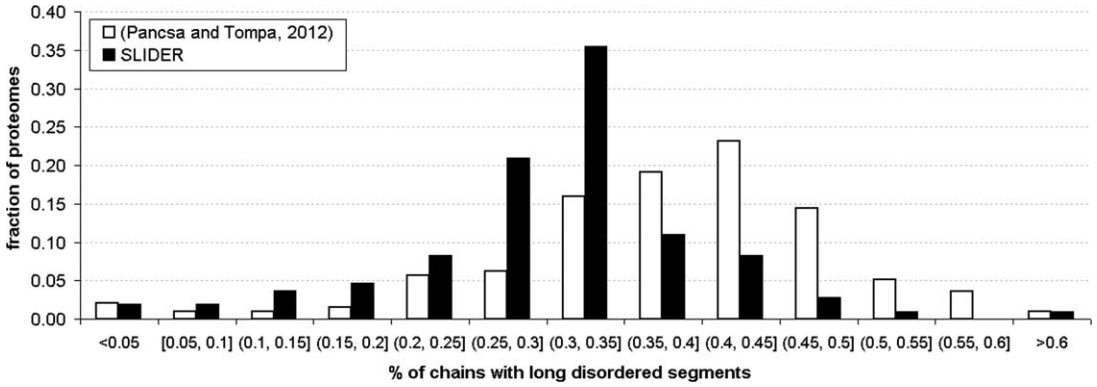

Figure 4

Comparison of distributions of fractions of chains with LDRs for 110 eukaryotic proteomes predicted with SLIDER and 194 proteomes predicted with IUPred in Ref. 34.

included 194 proteomes and which primarily focused on contrasting the intrinsic disorder between eukaryotes and prokaryotes.[34] The former work utilized predictions with an accurate VSL2B method, but it covered fewer species than we considered here, and the results were not broken down by different eukaryotic kingdoms/phyla, which would provide further insights. The latter study performed predictions with a less accurate, in the context of the prediction of proteins with LDRs, IUPred method and comprehensively analyzed ratios of chains with LDRs and ratios of AAs in LDRs for individual proteomes; we compare our results against this analysis. Another recent contribution looked into relations between the intrinsic disorder, proteome size, and organism complexity in 53 eukaryotes[74]; however, the authors did not analyze proteins with the long disordered segments.

Similar to works by Xue *et al.*[4] and Pancsa and Tompa[34] we quantified the overall abundance of proteins with LDRs and studied the relation of the presence of LDRs with the underlying chain length. The overall fraction of proteins with LDRs over the considered 110 eukaryotes is estimated to be 30.3%, which is similar to the estimate of 33.0% given by Ward *et al.*[3] and lower than 38.0% that was shown by Pancsa and Tompa.[34] Figure 4 gives side-by-side comparison of the distributions of fractions of proteins with LDRs across eukaryotic proteomes between the results by Pancsa and Tompa[34] and the results that we obtained using SLIDER. The two distributions have similar shape, which demonstrates that majority of the 110 proteomes have fairly substantial fractions of chains with LDRs, i.e., 25–40% of a given proteome computed with SLIDER and 30–50% computed with IUPred. We found relatively few proteomes on both tails of the distributions including those that are depleted (below 10%) or highly enriched (above 50%) in these proteins. The main difference is that our estimates are generally more conservative, which resulted in the shift to the left. Figure 5 summarizes the relation between the fractions of chains with LDRs and the average chain length per proteome for various eukaryotic phyla/kingdoms. Our results confirm the results from a smaller scale study by Xue *et al.*,[4] where the fractions of proteins with LDRs were shown to positively correlate with the chain length in the eukaryotic species; the corresponding Pearson Correlation Coefficient (PCC) equals 0.80 over all eukaryotes and ranges between 0.79 and 0.91 over various phyla/kingdoms. This relation is summarized in Figure 6 over the considered 1.9 million eukaryotic proteins, where the fraction of proteins with

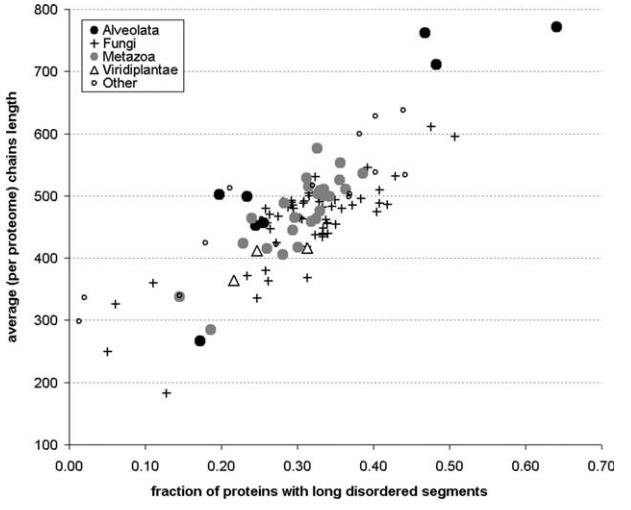

Figure 5

Relation between fraction (per proteome) of chains with LDRs (*x*-axis) and average (per proteome) chains size (*y*-axis) for the considered 110 eukaryotic proteomes. Markers types represent phyla/kingdoms: alveolata, fungi, metazoa, viridiplantae, and others that include amoebozoa, choanoflagellida, cryptophyta, diplomonadida, euglenozoa. parabasalia, and stramenopiles.

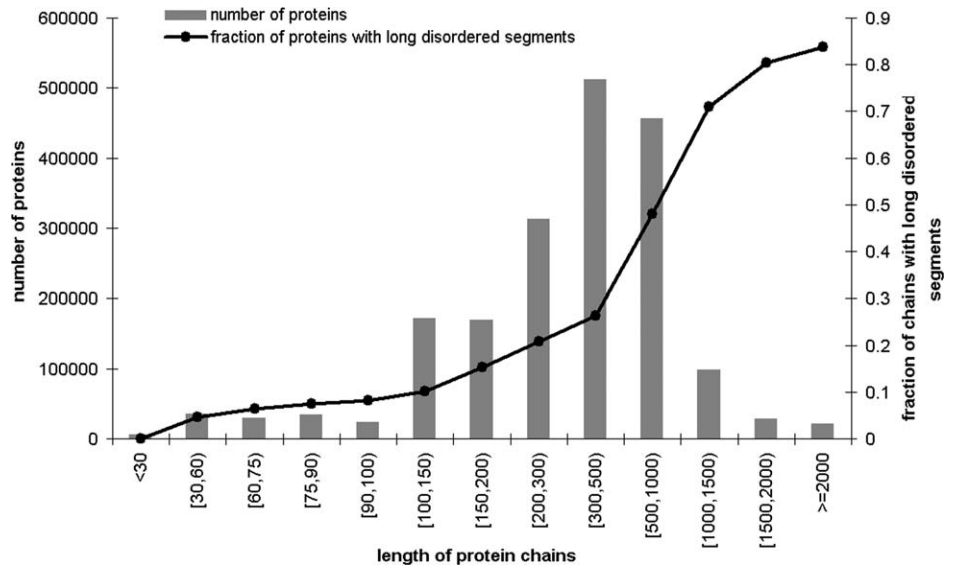

Figure 6

Relation between the fraction of chains with LDRs (black line) and chain length (*x*-axis) over the considered 110 eukaryotes. The gray bars show the corresponding number of chains for a given range of chain length.

LDRs grows monotonically with the chain size. Figure 6 shows that over 70% of large proteins that are composed of over 1000 AAs have LDRs. Our results that focus on eukaryotes are analogous to the results that were obtained based on proteins collected from the SwissProt databank.[75] Overall, we conclude that proteomes with bigger proteins are more likely to have more proteins with long disordered segments. An extreme example is the proteome of *Toxoplasma gondii*, which is located in the top right corner in Figure 5. This host-changing parasite has fairly large protein chains, and prior studies have suggested that the disorder in such parasitic organisms could be utilized to implement complex life cycle related to their ability to adapt to a wide range of environments.[34,76] The other species that include over 40% of chains with LDRs include two parasites from Plasmodium and another from Leishmania, several fungi, such as Ustilago, Trichophyton, Sporisorium, Filobasidiella, Neurospora, and Lodderomyces and a small algae Ectocarpus. Figure 5 also shows that fungi and alveolates are characterized by the widest spread of the content of chains with LDRs, spanning between 5 and 51% and between 17 and 64%, respectively. On the other hand, animals and plants have relatively small intra-species differences in the content of these chains, e.g., for the animals the spread is between 14 and 39%.

Functional characterization of chains with LDRs in eukaryotes

The prior characterizations of the cellular functions carried by proteins with LDRs were done only on a small scale. The earliest study considered only the *S. cerevisiae* proteome,[3] which was followed by an analysis of human proteome.[33] Several functions, such as transcription factor activity, DNA and protein binding, RNA metabolism, kinase signaling, and phosphorylation were found to be enriched in proteins with LDRs across these two studies.[33] We also note a couple of related studies, including analysis of a relatively small set of proteins collected from the SwissProt databank[75] and investigation of cellular localizations of proteins with LDRs in the human proteome.[35] The latter contribution found that proteins with long disordered segments are preferentially localized in nucleus, various membranes, cytoplasm, and endoplasmic reticulum.[35]

Our study is the first to investigate functions that are enriched in proteins with LDRs for a comprehensive set of eukaryotes. We utilize the prediction by SLIDER and the biological process and molecular function terms from Gene Ontology (GO),[77] which are linked to the proteins from the considered 110 eukaryotes that we collected from the UniProt. We discarded terms/annotations that are linked to <1000 chains over the 110 proteomes to assure that our analysis is statistically sound. The enrichment was calculated using a statistical test of significance based on the procedure introduced in.[78] Specifically, for a given annotation, we randomly selected half of proteins with this annotation and calculated their fraction that have LDRs. This fraction was compared with the corresponding fraction computed for the same number of size-matched chains (protein length was matched to be $\pm 10\%$) drawn at random from the entire set of 110 eukaryotes. The

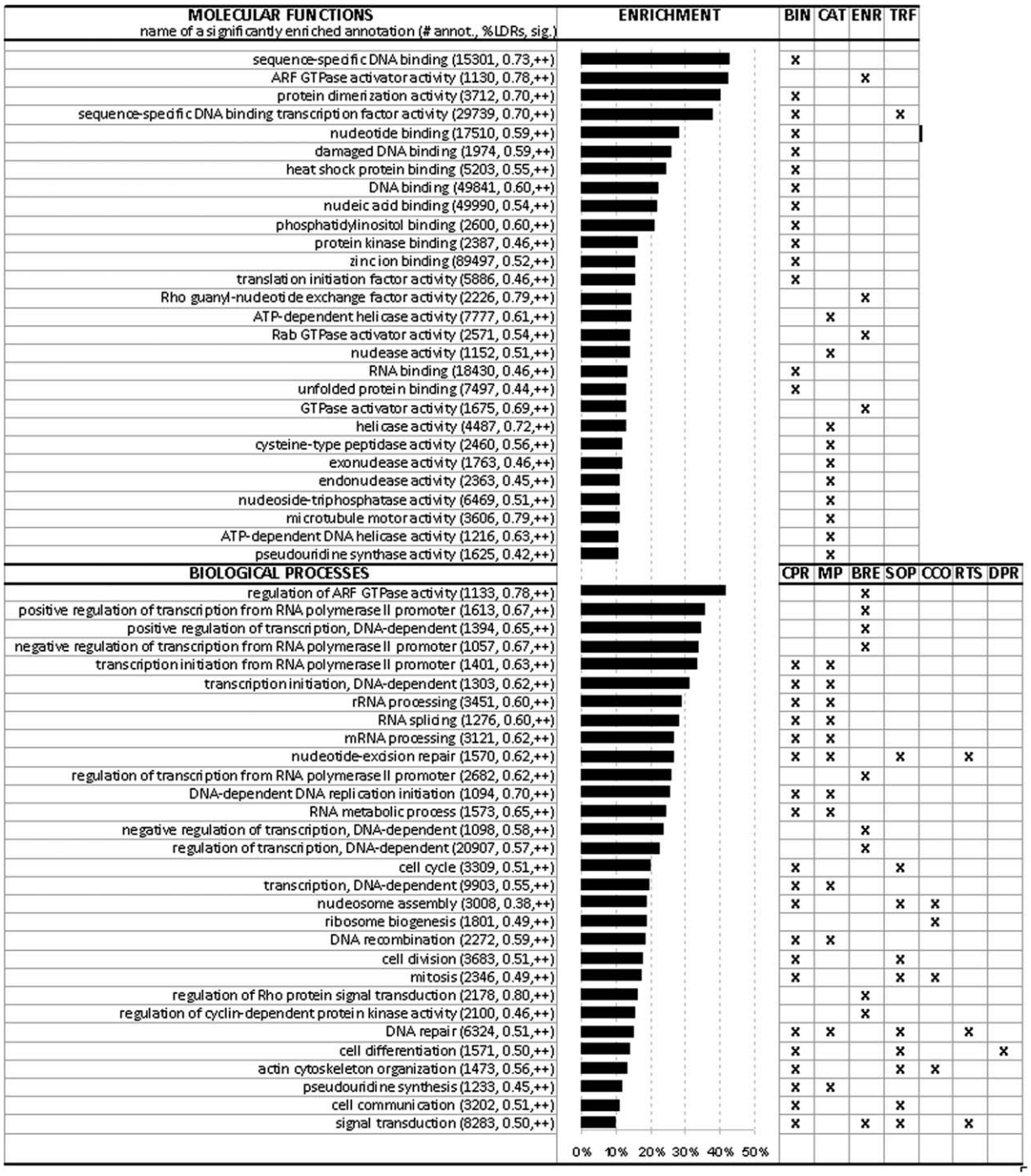

Figure 7

Molecular functions (top of the figure) and biological processes (bottom of the figure) that are significantly enriched in eukaryotic proteins with LDRs. The first column gives all significant functions/processes including their name, the number of corresponding annotated proteins (# annot.), fraction of these chains that have long disordered segment(s) (%LDRs), and significance of the enrichment (sig.). The significance is denoted with "+" and "++", which indicate that the *P*-value is <0.05 and <0.001, respectively. The functions/processes are sorted by the values of the difference. The horizontal bars in the second column shows the magnitude of the enrichment (difference in the fraction of proteins with LDRs between proteins annotated with a given functions and the baseline in the entire eukaryotic complete proteome). The remaining columns show the major types of molecular functions and biological processes and their association with the GO annotations; for the molecular functions they include binding (BIN), catalytic activity (CAT); enzyme regulator activity (ENR), and nucleic acid binding transcription factor activity (TRF), for the biological processes they include cellular process (CPR), metabolic process (MPR), biological regulation (BRE), single-organism process (SOP), cellular component organization or biogenesis (CCO), response to stimulus (RTS), and developmental process (DPR).

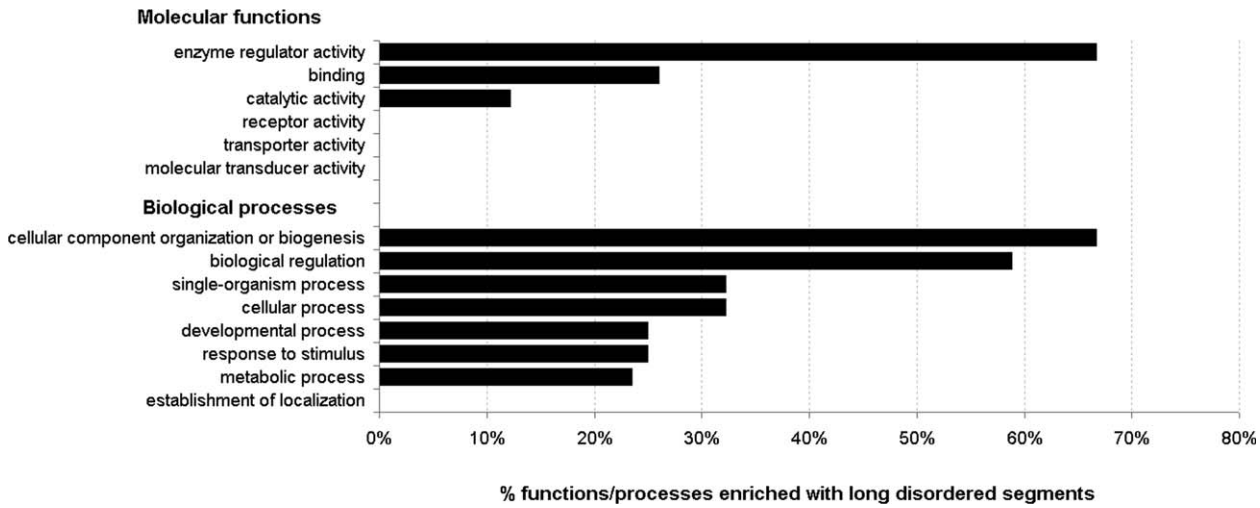

Figure 8

Major types of molecular functions (top of the figure) and biological processes (bottom of the figure) that are significantly enriched in eukaryotic proteins with LDRs. The specific functions/processes were aggregated to the second level in the GO ontology, which is shown on the left. Black bars show fractions of the considered annotations in a given second-level category that were found to be significantly enriched in chains with LDRs.

matching of the size was done to accommodate for the relation discussed in the "Abundance of chains with LDRs in eukaryotes" section. These calculations are repeated ten times and we assessed significance of the differences in the corresponding ten pairs of fractions. If these measurements follow normal distribution, as tested using Shapiro-Wilk test at the 0.05 significance, then we used paired *t*-test; otherwise we used the Wilcoxon rank sum test. A given annotation is assumed to be significantly enriched in proteins with LDRs if the resulting *P*-value <0.05. Furthermore, we only considered annotations where the increase in the fraction of chains with LDRs had sufficiently large magnitude, i.e., the increase must be at least 10% compared to the fraction of proteins with LDRs across all considered eukaryotic proteomes.

Figure 7 shows a relatively high number of cellular functions that are significantly enriched in proteins with LDRs. Some of these functions are associated with protein sets where majority of chains have LDRs. For instance, we identified eight molecular functions and five biological processes where over two-thirds of the corresponding chains have LDRs. Some of the enriched functions that we found overlap with the results from the previous studies,[3,33,75] including various DNA and RNA binding events that facilitate transcription, DNA repair, DNA replication, DNA recombination, and RNA metabolism; signal transduction via the GTPases and protein kinases; rRNA and mRNA processing; cell cycle processes; intracellular transport; metal ion binding; ribosome biogenesis and nucleosome assembly that was recently investigated by Peng *et al.*[79] Our results also point to the involvement of proteins with LDRs in some other functions and processes, including cell differentiation, cell division, and RNA splicing, which corroborates with the results obtained using SwissProt,[32,75] translation, protein dimerization, binding with unfolded proteins, several catalytic activities associated with cysteine peptidases, pseudouridine synthases, endonucleases and exonucleases, and involvement in activity of heat shock proteins, which was recently studied by Reichmann *et al.*[80] All considered individual functions and processes, including those that are and are not significantly enriched, were aggregated to the second level of GO to summarize our findings (see Fig. 8.) This figure lists major types of functions and processes that are enriched or depleted in proteins with LDRs and quantifies levels of their coverage; only the processes and functions that have enough data to provide a statistically sound evaluation are included. The proteins with LDRs were found to be functionally involved in about 26% of the considered binding annotations, in majority of catalytic regulation activities, and also to a smaller extent in the catalysis. These proteins are predominantly involved in processes related to cellular component organization, biogenesis, and biological regulation. The proteins with LDRs were also found to be enriched in about 24–32% of metabolic, developmental, and cellular processes, and in 25% of processes that implement responses to stimuli. On the other hand, Figure 8 also gives the major processes and functions that are less likely to utilize proteins with LDRs. They include establishment of cellular localization, transport of molecules, and transducer and receptor activities.

CONCLUSIONS

We have developed an accurate and fast method, called SLIDER, which predicts whether a given sequence has long disordered regions. A webserver that implements our predictor is available at http://biomine.ece.ualberta.ca/SLIDER/. Empirical tests show that SLIDER offers competitive predictive performance coupled with low computational cost, which allows for high-throughput, genome-scale applications. The strong predictive performance stems from our design that utilizes a carefully chosen set of custom-designed numerical features that quantify information extracted from selected physicochemical properties of amino acids, sequence complexity, and amino acid composition. Our method predicts an average size eukaryotic proteome in well under half an hour on a modern desktop computer and provides a 16-fold speedup compared to the best currently available approach.

SLIDER was used to perform large-scale investigation of occurrence and functional roles of proteins with LDRs in 110 eukaryotic proteomes. Our results are in agreement with prior studies that analyzed the abundance.[3,4,34] We showed that eukaryotes have substantial amounts of chains with LDRs, with the average of 30.3% proteins with LDRs and majority of proteomes having between 25 and 40%. We also demonstrated that proteomes that have larger proteins are more likely to have more proteins with LDRs. Such proteomes are characteristic to certain parasites and fungal species. Our first-of-its-kind large-scale analysis of the functional roles includes both confirmatory and novel results. Similar to studies that investigated these functional roles for yeast[3] and human[33] proteomes, we showed that chains with LDRs are enriched in transcription, DNA repair, replication, and recombination, RNA metabolism, signal transduction, rRNA and mRNA processing, metal ion binding, ribosome biogenesis and nucleosome assembly. Our analysis also reveals that proteins with LDRs are also involved in cell differentiation and division, RNA splicing, translation, protein dimerization, binding with unfolded proteins, and some catalytic activities.

ACKNOWLEDGEMENTS

This work was supported by the Alberta Innovates Graduate Student Scholarship in Omics to Z.P, the Dissertation fellowship awarded by University of Alberta to M.J.M., and the Discovery grant awarded by the Natural Sciences and Engineering Research Council of Canada to L.K.

REFERENCES

1. Dunker AK, Lawson JD, Brown CJ, Williams RM, Romero P, Oh JS, Oldfield CJ, Campen AM, Ratliff CM, Hipps KW, Ausio J, Nissen MS, Reeves R, Kang C, Kissinger CR, Bailey RW, Griswold MD, Chiu W, Garner EC, Obradovic Z. Intrinsically disordered protein. J Mol Graph Model 2001;19:26–59.
2. Dunker AK, Obradovic Z, Romero P, Garner EC, Brown CJ. Intrinsic protein disorder in complete genomes. Genome Inform Ser Workshop 2000;11:161–171.
3. Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J Mol Biol 2004;337:635–645.
4. Xue B, Dunker AK, Uversky VN. Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life. J Biomol Struct Dynam 2012; 30:137–149.
5. Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradovic Z. Intrinsic disorder and protein function. Biochemistry 2002;41: 6573–6582.
6. Dunker AK, Silman I, Uversky VN, Sussman JL. Function and structure of inherently disordered proteins. Curr Opin Struct Biol 2008;18:756–764.
7. Dyson HJ, Wright PE. Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Bio 2005;6:197–208.
8. Wright PE, Dyson HJ. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. J Mol Biol. 1999; 293: 321–331.
9. Tompa P, Fuxreiter M. Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions. Trend Biochem Sci 2008;33:2–8.
10. Tompa P, Fuxreiter M, Oldfield CJ, Simon I, Dunker AK, Uversky VN. Close encounters of the third kind: disordered domains and the interactions of proteins. Bioessays 2009;31:328–335.
11. Uversky VN, Oldfield CJ, Dunker AK. Intrinsically disordered proteins in human diseases: introducing the D(2) concept. Annu Rev Biophys 2008;37:215–246.
12. Uversky VN, Oldfield CJ, Midic U, Xie H, Xue B, Vucetic S, Iakoucheva LM, Obradovic Z, Dunker AK. Unfoldomics of human diseases: linking protein intrinsic disorder with diseases. BMC Genomics 2009;10:S7.
13. Xue B, Mizianty MJ, Kurgan L, Uversky VN. Protein intrinsic disorder as a flexible armor and a weapon of HIV-1. Cell Mol Life Sci 2012;69:1211–1259.
14. Sickmeier M, Hamilton JA, LeGall T, Vacic V, Cortese MS, Tantos A, Szabo B, Tompa P, Chen J, Uversky VN, Obradovic Z, Dunker AK. DisProt: the database of disordered proteins. Nucleic Acids Res 2007;35:D786–D793.
15. Dosztanyi Z, Meszaros B, Simon I. Bioinformatical approaches to characterize intrinsically disordered/unstructured proteins. Brief Bioinform 2010;11:225–243.
16. Liu JF, Tan HP, Rost B. Loopy proteins appear conserved in evolution. J Mol Biol 2002;322:53–64.
17. Uversky VN, Gillespie JR, Fink AL. Why are "natively unfolded" proteins unstructured under physiologic conditions? Proteins 2000; 41:415–427.
18. He B, Wang KJ, Liu YL, Xue B, Uversky VN, Dunker AK. Predicting intrinsic disorder in proteins: an overview. Cell Res 2009;19: 929–949.
19. Peng ZL, Kurgan L. Comprehensive comparative assessment of in-silico predictors of disordered regions. Curr Protein Pept Sci 2012; 13:6–18.
20. Deng X, Eickholt J, Cheng JL. A comprehensive overview of computational protein disorder prediction methods. Mol Biosyst 2012;8: 114–121.
21. Bordoli L, Kiefer F, Schwede T. Assessment of disorder predictions in CASP7. Proteins 2007;69:129–136.
22. Jin YM, Dunbrack RL. Assessment of disorder predictions in CASP6. Proteins 2005;61:167–175.
23. Melamud E, Moult J. Evaluation of disorder predictions in CASP5. Proteins 2003;53:561–565.

24. Noivirt-Brik O, Prilusky J, Sussman JL. Assessment of disorder predictions in CASP8. Proteins 2009;77: 210–216.
25. Monastyrskyy B, Fidelis K, Moult J, Tramontano A, Kryshtafovych A. Evaluation of disorder predictions in CASP9. Proteins 2011;79: 107–118.
26. Mizianty MJ, Zhang T, Xue B, Zhou Y, Dunker AK, Uversky VN, Kurgan L. In-silico prediction of disorder content using hybrid sequence representation. BMC Bioinformatics 2011;12:245.
27. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25: 3389–3402.
28. Walsh I, Martin AJM, Di Domenico T, Tosatto SCE. ESpritz: accurate and fast prediction of protein disorder. Bioinformatics 2012;28: 503–509.
29. Dosztanyi Z, Csizmok V, Tompa P, Simon I. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. J Mol Biol 2005;347:827–839.
30. Dosztanyi Z, Csizmok V, Tompa P, Simon I. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics 2005;21: 3433–3434.
31. Peng K, Radivojac P, Vucetic S, Dunker AK, Obradovic Z. Length-dependent prediction of protein intrinsic disorder. BMC Bioinformatics 2006;7:208.
32. Tompa P. Intrinsically disordered proteins: a 10-year recap. Trends Biochem Sci 2012;37:509–516.
33. Lobley A, Swindells MB, Orengo CA, Jones DT. Inferring function using patterns of native disorder in proteins. Plos Comput Biol 2007;3:1567–1579.
34. Pancsa R, Tompa P. Structural disorder in eukaryotes. Plos One 2012;7:e34687.
35. Pentony MM, Jones DT. Modularity of intrinsic disorder in the human proteome. Proteins 2010;78:212–221.
36. Chen JW, Romero P, Uversky VN, Dunker AK. Conservation of intrinsic disorder in protein domains and families: I. A database of conserved predicted disordered regions. J Proteom Res 2006;5: 879–887.
37. Xue B, Li L, Meroueh SO, Uversky VN, Dunker AK. Analysis of structured and intrinsically disordered regions of transmembrane proteins. Mol Biosyst 2009;5:1688–1702.
38. Korneta I, Bujnicki JM. Intrinsic disorder in the human spliceosomal proteome. Plos Comput Biol 2012;8:e1002641.
39. Tompa P, Fuxreiter M, Oldfield CJ, Simon I, Dunker AK, Uversky VN. Close encounters of the third kind: disordered domains and the interactions of proteins. Bioessays 2009;31:328–335.
40. Slabinski L, Jaroszewski L, Rodrigues AP, Rychlewski L, Wilson IA, Lesley SA, Godzik A. The challenge of protein structure determination—lessons from structural genomics. Prot Sci 2007;16: 2472–2482.
41. Mizianty MJ, Kurgan L. Sequence-based prediction of protein crystallization, purification and production propensity. Bioinformatics 2011;27:i24–i33.
42. Oldfield CJ, Xue B, Van YY, Ulrich EL, Markley JL, Dunker AK, Uversky VN. Utilization of protein intrinsic disorder knowledge in structural proteomics. Biochim et Biophys Acta 2013;1834:487–498.
43. Mizianty MJ, Stach W, Chen K, Kedarisetti KD, Disfani FM, Kurgan L. Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources. Bioinformatics 2010;26:i489–i496.
44. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. Nucleic Acids Res 2000;28:235–242.
45. Sirota FL, Ooi HS, Gattermayer T, Schneider G, Eisenhaber F, Maurer-Stroh S. Parameterization of disorder predictors for large-scale applications requiring high specificity by using an extended benchmark dataset. BMC Genomics 2010;11:S15.
46. Walsh I, Martin AJ, Di Domenico T, Vullo A, Pollastri G, Tosatto SC. CSpritz: accurate prediction of protein disorder segments with annotation for homology, secondary structure and linear motifs. Nucleic Acids Res 2011;39:W190–W196.
47. Walsh I, Martin AJ, Di Domenico T, Tosatto SC. ESpritz: accurate and fast prediction of protein disorder. Bioinformatics 2012;28: 503–509.
48. Zhang T, Faraggi E, Xue B, Dunker AK, Uversky VN, Zhou YQ. SPINE-D: accurate prediction of short and long disordered regions by a single neural-network based method. J Biomol Struct Dynam 2012;29:799–813.
49. Peng Z, Kurgan L. On the complementarity of the consensus-based disorder prediction. Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing, 2012; pp. 176–187.
50. Consortium U. Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic Acids Res 2012;40:D71–D75.
51. Shapiro SS, Wilk MB. An analysis of variance test for normality (complete samples). Biometrika 1965;52:591–611.
52. Wilcoxon F. Individual comparisons by ranking methods. Biometric Bull 1945;1:80–83.
53. Peng K, Radivojac P, Vucetic S, Dunker AK, Obradovic Z. Length-dependent prediction of protein intrinsic disorder. BMC Bioinformatics 2006;7:208.
54. Radivojac P, Obradovic Z, Smith DK, Zhu G, Vucetic S, Brown CJ, Lawson JD, Dunker AK. Protein flexibility and intrinsic disorder. Prot Sci 2004;13:71–80.
55. Kawashima S, Ogata H, Kanehisa M. AA index: amino acid index database. Nucleic Acids Res 1999;27:368–369.
56. Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M. AA index: amino acid index database, progress report 2008. Nucleic Acids Res 2008;36: D202–D205.
57. Romero P, Obradovic Z, Li X, Garner EC, Brown CJ, Dunker AK. Sequence complexity of disordered protein. Proteins 2001;42: 38–48.
58. Vacic V, Uversky VN, Dunker AK, Lonardi S. Composition Profiler: a tool for discovery and visualization of amino acid composition differences. BMC Bioinformatics 2007;8:211.
59. Williams RM, Obradovi Z, Mathura V, Braun W, Garner EC, Young J, Takayama S, Brown CJ, Dunker AK. The protein non-folding problem: amino acid determinants of intrinsic order and disorder. Pacific Symposium on Biocomputing, 2001; pp. 89–100.
60. Wootton JC. Non-globular domains in protein sequences: automated segmentation using complexity measures. Comput Chem 1994;18:269–285.
61. Wootton JC, Federhen S. Analysis of compositionally biased regions in sequence databases. Meth Enzymol 1996;266:554–571.
62. Tate RF. Correlation between a discrete and a continuous variable – point-biserial correlation. Ann Math Stat 1954;25:603–607.
63. Deng X, Eickholt J, Cheng J. PreDisorder: ab initio sequence-based prediction of protein disordered regions. BMC Bioinformatics 2009; 10:436.
64. Ishida T, Kinoshita K. PrDOS: prediction of disordered protein regions from amino acid sequence. Nucleic Acids Res 2007;35: W460–W464.
65. Yang ZR, Thomson R, McNeil P, Esnouf RM. RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. Bioinformatics 2005;21: 3369–3376.
66. McGuffin LJ. Intrinsic disorder prediction from the analysis of multiple protein fold recognition models. Bioinformatics 2008;24: 1798–1804.
67. Xue B, Dunbrack RL, Williams RW, Dunker AK, Uversky VN. PONDR-FIT: a meta-predictor of intrinsically disordered amino acids. Biochim Biophys Acta 2010;1804:996–1010.

68. Jones DT, Ward JJ. Prediction of disordered regions in proteins from position specific score matrices. Proteins 2003;53(Suppl 6): 573–578.
69. Schlessinger A, Liu J, Rost B. Natively unstructured loops differ from other loops. Plos Comput Biol 2007;3:e140.
70. Schlessinger A, Yachdav G, Rost B. PROFbval: predict flexible and rigid residues in proteins. Bioinformatics 2006;22:891–893.
71. Schlessinger A, Punta M, Yachdav G, Kajan L, Rost B. Improved disorder prediction by combination of orthogonal approaches. Plos One 2009;4:e4433.
72. Cheng JL, Sweredoski MJ, Baldi P. Accurate prediction of protein disordered regions by mining protein structure data. Data Min Knowl Disc 2005;11:213–222.
73. Schlessinger A, Punta M, Rost B. Natively unstructured regions in proteins identified from contact predictions. Bioinformatics 2007; 23:2376–2384.
74. Schad E, Tompa P, Hegyi H. The relationship between proteome size, structural disorder and organism complexity. Genome Biol 2011;12: R120.
75. Xie H, Vucetic S, Iakoucheva LM, Oldfield CJ, Dunker AK, Uversky VN, Obradovic Z. Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions. J Proteom Res 2007;6:1882–1898.
76. Mohan A, Sullivan WJ, Radivojac P, Dunker AK, Uversky VN. Intrinsic disorder in pathogenic and non-pathogenic microbes: discovering and analyzing the unfoldomes of early-branching eukaryotes. Mol Biosyst 2008;4:328–340.
77. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. Nature Genetics 2000;25:25–29.
78. Howell M, Green R, Killeen A, Wedderburn L, Picascio V, Alejandro A, Peng Z, Larina M, Xue B, Kurgan L, Uversky VN. Not that rigid midgets and not so flexible giants: on the abundance and roles of intrinsic disorder in short and long proteins. J Biolog Syst 2012;20: 471–511.
79. Peng Z, Mizianty MJ, Xue B, Kurgan L, Uversky VN. More than just tails: intrinsic disorder in histone proteins. Mol Biosyst 2012;8: 1886–1901.
80. Reichmann D, Xu Y, Cremers CM, Ilbert M, Mittelman R, Fitzgerald MC, Jakob U. Order out of disorder: working cycle of an intrinsically unfolded chaperone. Cell 2012;148:947–957.