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.

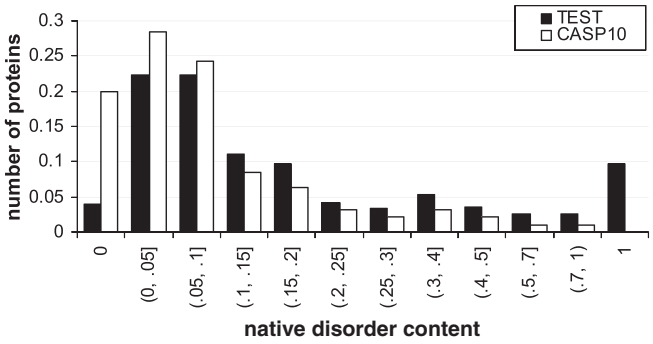

Fig. 1. Distribution of fraction of proteins (y-axis) in given intervals of the native disorder content for the TEST and CASP10 datasets. The x-axis shows the content binned to 0.05 wide intervals including values of 0 (fully structured proteins) and 1 (fully disordered proteins) on both ends.