FEBS Letters 583 (2009) 1469–1474

FEBS Letters

journal homepage: www.FEBSLetters.org

CDF it all: Consensus prediction of intrinsically disordered proteins based on various cumulative distribution functions

Bin Xue [a,b], Christopher J. Oldfield [a], A. Keith Dunker [a,b], Vladimir N. Uversky [a,b,c,*]

[a] Center for Computational Biology and Bioinformatics, Department of Biochemistry and Molecular Biology, Indiana University School of Medicine, 410 W. 10th Street, HS 5009, Indianapolis, IN 46202, USA

[b] Institute for Intrinsically Disordered Protein Research, Indiana University School of Medicine, Indianapolis, IN 46202, USA

[c] Institute for Biological Instrumentation, Russian Academy of Sciences, 142290 Pushchino, Moscow Region, Russia

ARTICLE INFO

Article history:
Received 28 January 2009
Revised 18 March 2009
Accepted 27 March 2009
Available online 5 April 2009

Edited by Robert B. Russell

Keywords:
Intrinsically disordered protein
Prediction
Accuracy
CDF

ABSTRACT

Many biologically active proteins are intrinsically disordered. A reasonable understanding of the disorder status of these proteins may be beneficial for better understanding of their structures and functions. The disorder contents of disordered proteins vary dramatically, with two extremes being fully ordered and fully disordered proteins. Often, it is necessary to perform a binary classification and classify a whole protein as ordered or disordered. Here, an improved error estimation technique was applied to develop the cumulative distribution function (CDF) algorithms for several established disorder predictors. A consensus binary predictor, based on the artificial neural networks, NN-CDF, was developed by using output of the individual CDFs. The consensus method outperforms the individual predictors by 4–5% in the averaged accuracy.

1. Introduction

The number of proteins lacking rigid 3D structures under physiological conditions in vitro yet fulfilling key biological functions is rapidly increasing [1–10]. These proteins are known as intrinsically disordered proteins (IDPs) among other names. They are highly abundant in nature [11–13], typically involved in signaling, recognition and regulation [7,8,14–18], and are strongly associated with human diseases [19]. IDPs typically possess highly dynamic structures in solution with high mobility at different timescales, and therefore such proteins almost never form crystals. Hence, the existence of these proteins represents a substantial challenge to the structural genomics initiative [20].

IDPs and IDRs differ from structured globular proteins and domains with regard to many attributes, including amino acid composition, sequence complexity, hydrophobicity, charge, flexibility, and type and rate of amino acid substitutions over evolutionary time [4,21–23]. Based on these differences between IDPs and ordered proteins, numerous disorder predictors have been developed (reviewed in Refs. [24–26]). Nearly all of the predictive tools developed so far provide disorder prediction on the per-residue basis; i.e., they give the likely disorder status of each amino acid residue. Often, in the analysis of a given dataset, it is useful to carry out a binary classification of whole proteins, indicating whether a protein is likely to fold or likely to remain unstructured. Such a classification is not a simple task, as the extent to which a sequence is ordered or disordered and the nature of disorder vary widely among proteins. In fact, the structural variability of IDPs is extremely high and native coils, native pre-molten globules, and native molten globules have been described in literature [4,9,10,14,16,18,27]. The protein can be completely unstructured or contain some elements of tertiary and/or secondary structure. In multi-domain proteins, domains might be connected by highly flexible linkers, and one or several domains might be completely disordered. Some proteins might have long disordered loops or tails. Because of this great variability, there is no strict boundary between ordered and intrinsically disordered proteins.

Two distinct binary classification methods have been reported previously [3,11,13]. One of these approaches uses charge-hydropathy plots (CH-plots), where ordered and disordered proteins are plotted in CH-space, and a linear boundary separates them [3]. The other method is based on predictor of natural disordered

Abbreviations: IDP, intrinsically disordered protein; IDR, intrinsically disordered region; CDF, cumulative distribution function; PONDR, predictor of natural disordered regions; PDD, partially disordered dataset

* Corresponding author. Address: Center for Computational Biology and Bioinformatics, Department of Biochemistry and Molecular Biology, Indiana University School of Medicine, 410 W. 10th Street, HS 5009, Indianapolis, IN 46202, USA. Fax: +1 317 278 9217.

E-mail address: vuversky@iupui.edu (V.N. Uversky).

0014-5793/$36.00 © 2009 Federation of European Biochemical Societies. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.febslet.2009.03.070

regions (PONDR®) VLXT [21,28], which predicts the order–disorder score for every residue in a protein. The cumulative distribution function (CDF) distinguishes ordered and disordered proteins based on the distribution of prediction scores [11,13]. A CDF curve gives the fraction of the outputs that are less than or equal to a given value. According to the CDF analysis, fully disordered proteins have very low percentage of residues with low predicted disorder scores, as the majority of their residues possess high predicted disorder scores. On the contrary, the majority of residues in ordered proteins are predicted to have low disorder scores. Hence, theoretically, all the fully disordered proteins should stay at the lower right half of the CDF plot, whereas all the fully ordered proteins should be located at the upper left half of this plot [11,13].

Due to the significant improvement in the prediction accuracy observed for several per-residue predictors, it was of interest to determine whether the CDF analysis based on these predictors would give improved binary classifications. An additional question was whether new methods can be used to optimize the CDF boundary line to achieve higher prediction accuracy. In this paper, the CDF method was developed for two other members of the PONDR® family of disorder predictors, VSL2 [29,30] and VL3 [31], for a simplified predictor based on the TOP-IDP scale [32], as well as for IUPred [33,34] and FoldIndex [35]. We also proposed a new method for optimizing the order–disorder boundary line in the CDF plots. Finally, a consensus method was elaborated by using a neural network based on CDF values from the outputs of the PONDR® VXLT, PONDR® VSL2, PONDR® VL3, TOP-IDP, IUPred, and FoldIndex, and this method appears to be more accurate than any of the methods based on individual predictors.

2. Materials and methods

2.1. Dataset construction

Four groups of datasets were used in this study. The first group included the 'original datasets' from Ref. [13]: (i) an ordered dataset of 105 wholly ordered proteins and (ii) a disordered dataset of 54 fully disordered proteins. These two datasets were used to take advantage of their high quality, and to provide an unambiguous comparison of the new methods developed in this paper with the previously developed method [13]. The second group was new fully ordered and fully disordered datasets. The new set of fully ordered proteins had 554 chains that were derived from the PDB database as of July 20, 2008 to include sequences of non-homologous single chain non-membrane proteins, which had no ligands, no disulfide bonds, and no missing residues, and which were characterized by unit cells with primitive space groups. The new dataset of fully disordered protein had 84 chains that were extracted from DisProt (release 4.5 of July 17, 2008) [36] to include non-homologous proteins without structured regions. Each of these new datasets was randomly and equally split into training and testing sets. The third group was the datasets of sequences for *Escherichia coli* K12, *Archaeoglobus fulgidus*, and *Methanobacterium thermoautotrophicum* generated from the UniProt database after removing all the fragments. The last group was a dataset that included 64 partially disordered proteins with less than 25% of sequence identity which were also extracted from PDB and had missing electron density for at least 30 residues, as in Ref. [13].

2.2. Individual disorder predictors and CDF

PONDR® VLXT [21,28] is composed of three neural networks, two for the termini of the sequence and one for internal region. The final output is an average over above three outs. The inputs of the neural networks are residue composition-related quantities. PONDR® VL3 [31] employs majority-voting over many neural networks which also take composition, complexity, and entropy as the inputs. PONDR® VSL2 [29,30] is built upon support vector machine with sequence composition, evolution information, and predicted secondary structure as the inputs. TOP-IDP [32] is a new amino acid scale developed to discriminate ordered and disordered residues with the highest accuracy. IUPred [33,34] applies a sequence-based pair-wise potential energy evaluated from globular proteins to distinguish disordered residues/proteins from the ordered ones. FoldIndex [35] takes the relative relation of net charges and normalized hydrophobicity scale which is originated from CH plot to partition ordered and disordered residues.

CDF analysis summarizes the per-residue predictions by plotting predicted disorder scores against their cumulative frequency, which allows ordered and disordered proteins to be distinguished based on the distribution of prediction scores [11,13]. At any given point on the CDF curve, the ordinate gives the proportion of residues with a disorder score less than or equal to the abscissa. To develop corresponding CDF algorithms, the outputs of all the above-mentioned predictors were unified to produce the per-residue disorder scores ranging from 0 (ordered) to 1 (disordered). In this way, CDF curves for various disorder predictors always began at the point (0, 0) and ended at the point (1, 1) because disorder predictions were defined only in the range [0, 1] with values less than 0.5 indicating a propensity for order and values greater than or equal to 0.5 indicating a propensity for disorder. As a result, fully ordered proteins yield convex curves because a high proportion of the prediction outputs are below 0.5, while fully disordered proteins typically yield concave curves because a high proportion of the prediction outputs are above 0.5. In practice, the range of prediction score (from 0 to 1) was divided into 20 bins [11,13]. It is expected therefore that there should be an approximately diagonal boundary line that could be used to separate the ordered and disordered proteins with an acceptable accuracy.

The original datasets were divided into training sets and testing sets. The boundary line for each CDF was optimized in the training set, and tested with the testing set. Bootstrap sampling of 1000 times was also applied to validate the confidence region of the accuracy.

A quantity termed CDF distance was also applied to assess whether the protein is ordered or disordered. The CDF distance is defined as:

$$\mathrm{dCDF} = \frac{\sum_{i=K_s}^{K_1}(\mathrm{CDF}_i - \mathrm{CDF}_i^0)}{K_1 - K_s + 1} \tag{1}$$

where dCDF is the averaged CDF distance of the protein from the CDF boundary line. K_s and K_e are the starting and ending bins of the CDF boundary line. CDF_i is the CDF value of *i*th bin, while CDF_i^0 is the value of CDF boundary at that bin.

2.3. Consensus prediction based on neural networks

By combining the CDFs based on PONDR® VLXT, PONDR® VSL2, PONDR® VL3, TopIDP, IUPred, and FoldIndex, a neural network-based consensus method of predicting the order/disorder status was developed. The neural network was fully connected with 20 inputs (three from the PONDR® VLXT-based CDF, four from the PONDR® VSL2-based CDF, three from the PONDR® VL3-based CDF, three from TopIDP-based CDF, four from IUPred-based CDF, and three from FoldIndex-based CDF), one hidden layer with 10 hidden units, and one output. A sigmoidal curve was used as the activation function at each node. Inputs from the CDF of each predictor were selected from the bins having the highest separating accuracies. The above-mentioned fully disordered and fully ordered datasets were randomly separated into eight groups with each group having one eighth of both the original training and test-