Structural bioinformatics

Prediction of disordered regions in proteins based on the meta approach

Takashi Ishida1,.* and Kengo Kinoshita1,2

1Institute of Medical Science, The University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo, 108-8639 and
2Structure and Function of Biomolecules, SORST JST, 4-1-8 Honcho, Kawaguchi, Saitama 332-0012, Japan

ABSTRACT

Motivation: Intrinsically disordered regions in proteins have no unique stable structures without their partner molecules, thus these regions sometimes prevent high-quality structure determination. Furthermore, proteins with disordered regions are often involved in important biological processes, and the disordered regions are considered to play important roles in molecular interactions. Therefore, identifying disordered regions is important to obtain high-resolution structural information and to understand the functional aspects of these proteins.

Results: We developed a new prediction method for disordered regions in proteins based on the meta approach and implemented a web-server for this prediction method named ‘metaPrDOS’. The method predicts the disorder tendency of each residue using support vector machines from the prediction results of the seven independent predictors. Evaluation of the meta approach was performed using the CASP7 prediction targets to avoid an overestimation due to the inclusion of proteins used in the training set of some component predictors. As a result, the meta approach achieved higher prediction accuracy than all methods participating in CASP7.

Availability: http://prdos.hgc.jp/meta/

Contact: t-ishida@hgc.jp

1 INTRODUCTION

Intrinsically disordered regions that have no stable structures without their partner molecules are often found in functional sites of proteins, especially eukaryotic proteins (Dunker et al., 2001; Ward et al., 2004). The functions of proteins with disordered regions are now believed to be quite varied, and such proteins play a critical role in the molecular-interaction network of the cell. For example, disordered regions are involved in transcription, translation and cell signaling (Dyson and Wright, 2005; Uversky et al., 2005), as well as in alternative splicing (Romero et al., 2006). Furthermore, the primary role of disordered regions is considered to be the recognition of other partner molecules, such as proteins, DNA, or RNA (Dunker et al., 2002; Dyson and Wright, 2002), because the flexibility of disordered regions may be more adaptable for subsequent interaction with multiple partners with high specificity but low affinity (Dunker et al., 2001).

Identification of disordered regions in proteins is important for the functional annotation of proteins and for high-throughput structural determination, because disordered regions often lead to difficulties in purification and crystallization (Oldfield et al., 2005). Usual methods to identify disordered regions experimentally are X-ray crystallography, NMR spectroscopy, circular dichroism spectroscopy and protein proteolysis (Wright and Dyson, 1999). However, it is almost impossible to experimentally determine all disordered regions encoded by genomes, and thus computational methods that predict disordered regions from amino acid sequences are necessary, and various prediction methods have been proposed (Ferron et al., 2006). Disordered regions tend to have particular physiochemical properties reflecting amino acid composition such as high-netcharge, low hydrophobicity, and/or low sequence complexity (Dunker et al., 2001). The simplest approach is to use these physiochemical features calculated from amino acid composition directly (Prilusky et al., 2005; Uversky, 2002). Other approaches consider the contact propensities and pairwise interaction energy (Dosztanyi et al., 2005a), or they incorporate evolutionary information in the form of sequence profiles (Jones and Ward, 2003). Some prediction methods also introduce additional information, such as predicted secondary structure (Ward et al., 2004), predicted accessible surface area (Cheng et al., 2005) or structural templates (Ishida and Kinoshita, 2007). In addition, disorder prediction is now one of the categories of the Critical Assessment of Techniques for Protein Structure Prediction (CASP) experiments (Melamud and Moult, 2003), which might promote the development of a new method for prediction of disordered regions. As a result, the number of prediction methods available through the internet has increased rapidly (Ferron et al., 2006), enabling to use the meta approach to predict disordered regions.

The meta, or consensus, approach is a method used to make a prediction by integrating the prediction results of several methods. The meta approach has already been used in the field of protein tertiary structure prediction (Bujnicki et al., 2001a; Ginalski et al., 2003; Lundstrom et al., 2001), and some critical experiments showed the improved performance of meta predictors when compared with the individual methods used in
the meta predictors (Bujnicki et al., 2001b; Fischer et al., 2001). The meta approach also has been applied to protein domain predictions and has shown better performance in that area as well (Saini and Fischer, 2005).

Here, we report a new method to predict disordered regions of proteins based on the meta approach, and its evaluation. Our method predicts disordered regions by integrating the results of seven different prediction methods. Assessing the performance of meta prediction is not straightforward because it is almost impossible to eliminate all proteins that may be related genetically to the proteins used in the training set of each component from the test sets, and inclusion of similar proteins in the test set will cause overestimation of prediction accuracy. Therefore, here we evaluated the performance of the meta approach by preparing the latest CASP7 (Bordoli et al., 2007) prediction targets as the test set, which enabled us to compare the prediction results with other methods used in CASP7.

2 METHODS

2.1 Meta prediction

Meta prediction comprises two main steps as shown in Figure 1. In the first step, an input sequence is submitted to each disorder predictor, and prediction results from all predictors are collected. In this study, we used seven predictors: PrDOS (Ishida and Kinoshita, 2007), DISOPRED2 (Ward et al., 2004), DisEMBL (Linding et al., 2003), DISPROT (VSL2P) (Peng et al., 2006), DISpro (Cheng et al., 2005), IUpred (Dosztanyi et al., 2005b) and POODLE-S (Shimizu et al., 2007). These predictors were selected according to their prediction accuracy and availability. Each predictor will perform its own prediction for each residue, and the result is obtained as a disorder tendency (a numerical value). In the second step, the meta predictor integrates the prediction results and determines the disorder tendency for each residue. Thus, the dimension of the input vector for meta predictor corresponds to the number of component predictors. Because the prediction sensitivity and scaling method differ among component predictors, a simple meta approach such as using a consensus or averaging the results of component predictors would be insufficient in this case. Therefore, we adopted the support vector machine (SVM) (Vapnik, 1998) as the meta predictor in this study and employed the libSVM version 2.82 package (Fan et al., 2005) to implement the meta predictor. Because SVM is a binary classifier, the SVM outputs two-state prediction results. However, a decision value—the distance between each input vector and a decision plane—can be used to evaluate the reliability of the prediction. In general, the prediction with a higher decision value is considered as more reliable (Vapnik, 1998).

In our method, the decision value of the SVM is scaled from 0.0 to 1.0, and it is returned as a prediction result. The details of this scaling method are described in Section 2.5.

2.2 Training dataset

First, we constructed a non-redundant protein chain set from the Protein Data Bank (PDB) as of April 2006 (Berman et al., 2000), using the PISCES server (Wang and Dunbrack, 2005). The set was selected using the following criteria: determined by X-ray crystallography with resolution ≤2.5 Å and R-factor ≤0.25, sequence identities to each other ≤20%, and sequence length >50 residues. Chains including non-standard amino acids and chains with sequence identities >20% to chains used in the training of the PrDOS predictor (Ishida and Kinoshita, 2007) were excluded. Disordered regions of these proteins were identified as the missing residues according to the ‘REMARK 465’ lines in the header of each PDB entry. As a result, 486 chains with disordered regions were selected, which had 7368 disordered residues (5.9%) and 117,967 ordered residues (94.1%).

2.3 CASP7 set

To assess prediction performance, prediction targets in CASP7 were used as a test dataset, and were obtained from http://predictioncenter.org/download_area/CASP7/. The set contains 96 structures and 19,816 residues. The CASP7 committee provided the state of each residue, 18,627 ordered and 1,189 disordered residues thereby were obtained (Bordoli et al., 2007). We did not exclude sequence similarities between the training dataset and CASP7 set. However, the highest sequence similarity was under 40% and most of them distributed under 30%.

2.4 Evaluation measure

The number of ordered residues is far greater than that of disordered residues. Thus, the Q2 accuracy, a percentage of correctly predicted residues in a two-state prediction, is not suitable for this analysis because a method predicting all residues as ordered can easily achieve the highest Q2 accuracy. To overcome this difficulty, we used two different measures, the receiver-operator characteristics (ROC) curve (Zweig and Campbell, 1993) and the Matthews correlation coefficient (MCC) (Matthews, 1975), to evaluate the prediction accuracy. An ROC curve is a plot of sensitivity and specificity (or false positive rate = 1 – specificity), and shows the trade-off between sensitivity and specificity (Zweig and Campbell, 1993). Here, we regard a disordered residue as positive, thus the number of true positives (TP) is the number of residues defined as disordered and predicted as disordered with a given threshold to judge if the residue is disordered. Similarly, the number of true negatives (TN), of false negatives (FN), and of false positives (FP) are counted with the same threshold, and the specificity and sensitivity are calculated as TP/(TP + FN) and TN/(FP + TN), respectively. Then, an ROC curve is obtained by changing the threshold values from strict to loose. When the area under the ROC curve of a predictor is larger than the area of other ROC curves, the predictor is regarded as a better predictor. The area under an ROC curve will be regarded as the ROC score, in this article. The MCC was calculated as follows:

\[
MCC = \frac{(TP*TN) - (FN*FP)}{\sqrt{(TP + FP)*(TN + FN)*(TP + FN)*(TN + FP)}},
\]

and evaluates the balance between FP and TP.
2.5 Scaling method for final prediction values

The decision values of the SVM were scaled from 0.0 to 1.0 and used as prediction results for the meta predictor. The decision values are the signed distances from the decision plane to each input vector. A large absolute value indicates a reliable prediction and positive signs are assigned for more disorder-like vectors. With this definition of decision values, we can usually find a threshold, $T_{\text{max}}$, that makes all predicted residues truly in the disordered state (i.e., specificity = 1) in the cross validation tests. Similarly, a threshold value, $T_{\text{min}}$, can be found that achieves a sensitivity = 1, where all the disordered states are correctly predicted with many FP. Therefore, if we change the threshold value to judge the disorder state from $T_{\text{max}}$ to $T_{\text{min}}$, then we will be able to calculate the disorder tendency, $d$, in accordance with the desirable FP rate using the following formula:

$$
\begin{align*}
    d &= 1.0 & \text{if } v \geq T_{\text{max}} \\
    d &= 0.5 + ((v - T_{\text{FP}})/(T_{\text{max}} - T_{\text{FP}})) \times 0.5 & \text{if } T_{\text{FP}} \leq v < T_{\text{max}} \\
    d &= 0.5 - ((v - T_{\text{FP}})/(T_{\text{min}} - T_{\text{FP}})) \times 0.5 & \text{if } T_{\text{min}} < v < T_{\text{FP}} \\
    d &= 0 & \text{if } v < T_{\text{min}}
\end{align*}
$$

where $v$ is the decision value, and $T_{\text{FP}}$ is a threshold value giving the desired FP rate when $d = 0.5$. In this study and on our web server, we set the default value as 0.05 for the FP rate. The disorder tendency is based on the decision values of SVMs, and cannot produce statistical meaning. However, the disorder tendency represents the confidence of the prediction and shows a good correlation to the sensitivity of the prediction. Thus, this value is practically useful for the users.

3 RESULTS

3.1 Training of the meta predictor

To optimize the training parameters of the SVM method in the meta predictor, a 10-fold cross validation approach was used. In this approach, the chains in the test set were first randomly divided into 10 subsets, and 1 of these subsets was reserved for an evaluation, whereas the other 9 subsets were used as training for the meta predictor. Then, the evaluation using an ROC score was repeated 10 times using each subset independently as an evaluation set. Finally, the best meta predictor was selected according to its ROC score. It yielded an ROC score of 0.904 ($\pm 0.004$) and an MCC value of 0.526 ($\pm 0.009$) (Table 1). The values in parentheses give the 95% confidence intervals. The ROC score and MCC of individual component predictors for the training set of meta predictor were also calculated. The most accurate component predictor yielded an ROC score of 0.887 and an MCC value of 0.502, and the prediction accuracy of meta prediction was thereby superior to any individual predictor for both the ROC score and MCC.

3.2 Performance evaluation using the CASP7 set

In general, sequence redundancy between training and test sets should be eliminated in the assessment to prevent artificially high accuracy. However, it is almost impossible to eliminate this sequence redundancy using the meta approach as described. Thus, the prediction performance of meta predictor was evaluated using the CASP7 set to obtain a more reliable assessment. Figure 2 shows an ROC curve of the CASP7 set for the meta predictor and the four most successful groups in the benchmark: ’ISTZORAN’, ‘CBRC-DR’, ‘fais’ and ‘DISOPRED’. These data were obtained from http://predictioncenter.org/download_area/CASP7/predictions/DR283-386.tar.gz. The curve of the meta predictor showed higher prediction performance on the whole range of FP rates, and it yielded an ROC score of 0.877 ($\pm 0.007$) and an MCC value of 0.440 ($\pm 0.013$). These values were superior to those obtained with the other prediction methods we examined (Table 2). The values in parentheses are standard errors and were calculated according to the public assessment of CASP7 (Bordoli et al., 2007). We also performed Student’s $t$ test for MCC distribution of the meta approach and that of each prediction group to check the statistical significance of the advantage of meta approach. MCC distributions were generated by using bootstrap samplings (1000 times). As results, the meta approach showed significantly higher performance for all prediction groups (all $P$-values < 0.001).

3.3 Web-based interface for the meta method

A web server named metaPrDOS was constructed as an interface for our meta prediction method at http://prdos.hgc.jp/meta/. The metaPrDOS server is freely available for academic users. It requires a single amino acid sequence as an input.
input and an e-mail address to send the prediction result. When a sequence is submitted, the metaPrDOS server forwards the sequence information to external servers to obtain their prediction results, that will then be used as an input for the meta predictor. In our experience, occasionally some external servers did not reply to our forwarded query within an hour (predefined limit in the metaPrDOS server), because they may have been down temporarily or they may set a limit to handle requests in a single day from the same IP address. At that time, metaPrDOS will make a prediction using all available prediction results. For this purpose, we prepared trained meta predictors in advance for all possible combinations of predictors. When the number of available servers was limited, the performance of the meta approach deteriorated (see discussion for detail). Finally, the prediction result by metaPrDOS will be sent to the user via e-mail. To facilitate easy interpretation of the result, the e-mail also contains a URL of the result web page. The result web page shows the two-state prediction results with a given FP positive rate for the users convenience, and the disorder profile plot is also shown to facilitate intuitive interpretation of the results (Fig. 3).

4 DISCUSSION

Seven component predictors were used to construct the meta predictor presented in this work. Among them, some predictors show relatively low-prediction accuracy as shown in Table 1. Although our meta predictor outperforms each of the individual component predictors, and indeed it may be possible that some predictors negatively influence the results; thus, the current composition of our meta predictor may not be optimal. To analyze this possibility, we examined the relationship between the number of component predictors and their performance (Fig. 4).

First, the two best component predictors, PrDOS and DISpro, were selected to make a meta predictor, and then the predictors with the next highest performance as determined by ROC scores were added one by one to construct a series of meta predictors. As shown in Figure 4, even predictors with relatively low-prediction accuracy can improve the prediction performance of the meta predictors, and meta predictors with more component predictors were generally found to yield higher performance. However, more predictors do not always cause an improvement of the outcome. Actually, when a random predictor returning values ranging from −1 to 1, which are independent of inputs, were added as the component predictor, the prediction accuracy decreased (ROC score = 0.902). Therefore, each component predictor should be carefully selected, but it is unclear how to select component predictors before evaluating the constructed meta predictor. From the view point of machine learning theories, the meta-prediction approach is a derivative of ensemble learning such as bagging.
ACKNOWLEDGEMENTS

This work was partly supported by a Grant-in-Aid for Scientific Research on Priority Areas ‘Transportsome’ from the Ministry of Education, Culture, Sports and Technology of Japan, and by a Grant-in-aid from the Institute for Bioinformatics Research and Development, Japan Science and Technology Corporation (BIRD-JST) to K.K. Computation time was provided by the Super Computer System, Human Genome Center, Institute of Medical Science, The University of Tokyo.

Conflict of Interest: none declared.

REFERENCES


