FDserver: a web service for protein folding research

Bin-Guang Ma[1,2,*], Ling-Ling Chen[1] and Hong-Yu Zhang[1,*]

[1] Shandong Provincial Research Center for Bioinformatic Engineering and Technique, Center for Advanced Study, Shandong University of Technology, Zibo 255049, P. R. China and [2] College of Chemistry and Chemical Engineering, Suzhou University, Suzhou 215006, P. R. China

ABSTRACT

Summary: To facilitate the study of protein folding, we have developed a web service for protein folding rate and folding type prediction as well as for the calculation of a variety of topological parameters of protein structure, which is freely available to the community.

Availability: http://sdbi.sdut.edu.cn/FDserver

Contact: bgMa@sdut.edu.cn

1 INTRODUCTION

Protein folding is the process that a protein transforms from its denatured state to its native state with specific biological function. Folding rate is a measure of how fast this transformation process is while folding type indicates the kinetic behavior of this process (Jackson, 1998; Fersht and Daggett, 2002). It has been observed that proteins have a wide range of folding rates from within microseconds to almost hours (Ivankov and Finkelstein, 2004). The folding behaviors are also not unique but can be roughly classified as two sorts: two-state folding and multi-state folding. Protein folding is one of the central research fields of molecular biology.

To predict protein folding rate and folding type, many parameters have been proposed in the past decade. Till now, several web services have been established (Zhou and Zhou, 2002; Gromiha *et al.*, 2006; Capriotti and Casadio, 2007). However, the sequence-based folding type prediction has not been implemented yet. In our previous works, we studied systematically the relationship between a protein's folding rate/type and its sequence/structure determinants and proposed three predictors, CI, Cp and I_{FT} for folding rate and folding type prediction, respectively (Ma *et al.*, 2006; Ma *et al.*, 2007). Jointly with a newly developed indicator of protein folding rate I_{RT}, the present work will implement a sequence-based predictor and a structure-based predictor for protein folding rate and folding type prediction. Moreover, the present work will also provide a calculator for protein topological parameters that stemmed from the protein folding research. This server is designed intentionally to facilitate protein folding and related research.

2 DESCRIPTION OF FDSERVER

FDserver comprises three components: (i) CIpred, a sequence-based protein folding rate and folding type predictor; (ii) SBpred, a structure-based folding rate and folding type predictor; (iii) TPcalc, a calculator for the topological parameters of protein structure. The predictors are trained on a large dataset and the calculator can calculate nine topological parameters of protein structure.

2.1 Training Data

The training data are composed of 85 proteins, an ever largest dataset to our knowledge, 43 of which are two-state folders and the others are multi-state folders. Among these proteins, 62 natural ones with experimentally determined folding rates (Ivankov and Finkelstein, 2004) are used for training the folding rate predictors (Ma *et al.*, 2006), while the totality are used for training the folding type predictors (for the detailed list of the proteins, see Ma *et al.*, 2007). The training data are also provided on the FDserver web site.

2.2 CIpred: sequence-based folding rate and folding type predictor

Based on the findings in our previous work, the sequence-based folding rate prediction is accomplished by using Composition Index (CI) (Ma et al., 2006). On the training dataset, the correlation coefficients (R) between CI and the folding rates of two-state folders, multi-state folders and their mixture are 0.73, 0.70 and 0.73, respectively (jack-knife test), with average deviations all less than 1 unit (Ma *et al.*, 2006). The folding type prediction is performed by using the predictor Cp which is defined based on the amino acid composition of a protein sequence (Ma *et al.*, 2007). The folding type of a protein is indicated by the value of Cp, i.e., if $Cp > 0$, the folding type is predicted as multi-state; otherwise, two-state. On the training dataset, the accuracy (jack-knife test) of the folding type prediction is 81% (Ma *et al.*, 2007).

2.3 SBpred: structure-based folding rate and folding type predictor

A structure-based predictor called SBpred is also implemented which can predict the protein folding rate and folding type with relatively higher accuracy than that based only on sequence information. SBpred predicts protein folding type according to the value of a folding type indicator I_{FT} defined in our previous work (Ma *et al.*, 2007), while it predicts folding rate according to a newly defined indicator I_{RT} whose definition can be found on the "Help" page of this server. On the training dataset, by using I_{FT}, the prediction accuracy of folding type is 85% (jack-knife test) while by using I_{RT}, the correlation coefficients (R) between I_{RT} and the experimental folding rates are 0.80 for back-check and 0.76 for jack-knife test. See the "Help" page of this service for a detailed description.

2.4 TPcalc: a calculator for topological parameters of protein structure

TPcalc can calculate nine parameters that measure the topology of a protein structure. These parameters are collected from literature, including relative contact order (CO), absolute contact order (Abs_CO), chain topology parameter (CTP), total contact distance (TCD), long range order (LRO), fraction of local contacts (Flocal), long range contact order (LR_CO), clustering coefficient (CC) and absolute clustering coefficient (Abs_CC). The definition of these parameters and the corresponding publications can be found on the "Help" page of this service. For some parameters, thresholds exist

*To whom correspondence should be addressed.

for the sequence separation (called L_{cut}) and/or distance definition (called R_{cut}) between two contacting residues. These thresholds are given the default values as reported in the corresponding original publications. It is also allowed for users to set these cut-off values on the server before calculation is performed.

2.5 Implementation

FDserver is built on a LAMP (Linux + Apache + MySQL + PHP) architecture and the web interface is shown in Figure 1. Through the home page (Fig. 1a), users can access the three components (Fig. 1b, 1c, 1d) of this server. For the calculation of topological parameters, users can submit a protein structure to TPcalc and select the parameters interested to perform calculation. For prediction of protein folding rate and/or folding type, if the protein structure information is available, SBpred can be employed. It accepts a protein structure as input and output the predicted folding rate and/or folding type. If only sequence information is available, a protein sequence can be submitted to the CIpred server, and then the folding type of the protein should be chosen to make a folding rate prediction; one may also make a folding type prediction at first when the folding type of the protein is unknown. The prediction results will be returned on a new web page. There is also an optional e-mail service on this server for delivering results. If an email address is provided, it will also receive the predicted or calculated results.

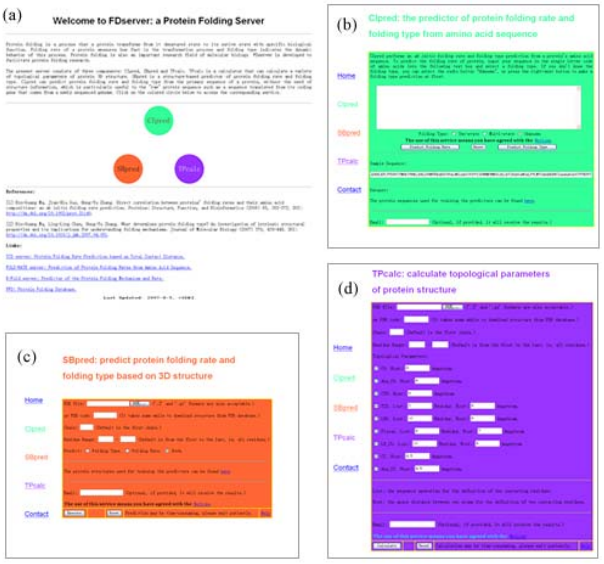

Fig. 1. The web interface of FDserver. (a) The home page; (b) the CIpred page; (c) the SBpred page; (d) the TPcalc page.

3 COMPARISON WITH AND LINKS TO OTHER SERVERS

Compared with other servers, FDserver is currently the only one that can predict protein folding rate and folding type simultaneously directly from the primary sequence of a protein and the prediction accuracy is comparable to the other severs, which is particularly useful to new sequenced proteins that lack 3D structure information. Moreover, the present server provides a calculator for a variety of topological parameters which has never been served before. The present server also maintains links to the other servers and to the protein folding database (Fulton *et al.*, 2005). Therefore, it can be treated as a catalog for the resources of protein folding research and brings convenience to the users.

4 LIMITATIONS AND POSSIBLE IMPROVEMENTS

The present approach is a statistical approach and thus inevitably has dataset dependency as all the statistical approaches do. Although the training dataset used in this work is unprecedentedly large, the amount of the currently available data is still small considering the whole protein universe. In addition, the quality of the data is also not completely satisfying because of the discrepant experimental conditions of data acquisition. The appearance of the so-called “standard” set of experimental conditions for the kinetic study of protein folding may provide us with an ideal training dataset in the future (Maxwell *et al.*, 2005).

ACKNOWLEDGEMENTS

This work was supported by the National Basic Research Program of China (2003CB114400) and the National Natural Science Foundation of China (30600119 and 30570383).

REFERENCES

Capriotti, E. and Casadio, R. (2007) K-Fold: a tool for the prediction of the protein folding kinetic order and rate. *Bioinformatics*, **23**, 385-386.

Fersht, A.R. and Daggett, V. (2002) Protein folding and unfolding at atomic resolution. *Cell*, **108**, 573-582.

Fulton, K.F., Devlin, G.L., Jodun, R.A., Silvestri, L., Bottomley, S.P., Fersht, A.R. and Buckle, A.M. (2005) PFD: a database for the investigation of protein folding kinetics and stability, *Nucleic Acids Res.*, **33**, D279-283.

Gromiha,M.M. et al. (2006) FOLD-RATE: prediction of protein folding rates from amino acid sequence. *Nucleic Acids Res*, **34**, W70-74.

Ivankov, D.N. and Finkelstein, A.V. (2004) Prediction of protein folding rates from the amino acid sequence-predicted secondary structure. *Proc. Natl. Acad. Sci. USA*, **101**, 8942-8944.

Jackson, S.E. (1998) How do small single-domain proteins fold? *Fold Des*, **3**, R81-91.

Ma, B.G. et al. (2006) Direct correlation between proteins' folding rates and their amino acid compositions: An ab initio folding rate prediction. *Proteins*, **65**, 362-372.

Ma, B.G. et al. (2007) What determines protein folding type? An investigation of intrinsic structural properties and its implications for understanding folding mechanisms. *J. Mol. Biol.* **370**, 439-448.

Maxwell, K.L. et al. (2005) Protein folding: Defining a "standard" set of experimental conditions and a preliminary kinetic data set of two-state proteins. *Protein Sci*, **14**, 602-616.

Zhou, H. and Zhou, Y. (2002) Folding rate prediction using total contact distance. *Biophys J*, **82**, 458-463.