BIOINFORMATICS APPLICATIONS NOTE Vol. 23 no. 3 2007, pages 385–386
doi:10.1093/bioinformatics/btl610

Structural bioinformatics

K-Fold: a tool for the prediction of the protein folding kinetic order and rate

E. Capriotti*,† and R. Casadio

Biocomputing Group, CIRB/Department of Biology, University of Bologna, via Irnerio 42, 40126 Bologna, Italy

Received on September 22, 2006; revised on November 15, 2006; accepted on November 24, 2006
Advance Access publication November 30, 2006
Associate Editor: Anna Tramontano

ABSTRACT

Summary: K-Fold is a tool for the automatic prediction of the protein folding kinetic order and rate. The tool is based on a support vector machine (SVM) that was trained on a data set of 63 proteins, whose 3D structure and folding mechanism are known from experiments already described in the literature. The method predicts whether a protein of known atomic structure folds according to a two-state or a multi-state kinetics and correctly classifies 81% of the folding mechanisms when tested over the training set of the 63 proteins. It also predicts as a further option the logarithm of the folding rate. To the best of our knowledge, the tool discriminates for the first time whether a protein is characterized by a two state or a multiple state kinetics, during the folding process, and concomitantly estimates also the value of the constant rate of the process. When used to predict the logarithm of the folding rate, K-Fold scores with a correlation value to the experimental data of 0.74 (with a SE of 1.2).

Availability: http://gpcr.biocomp.unibo.it/cgi/predictors/K-Fold/K-Fold.cgi

Contact: emidio@biocomp.unibo.it

Supplementary information: http://gpcr.biocomp.unibo.it/~emidio/K-Fold/K-Fold_help.html

1 INTRODUCTION

Protein folding is one of most relevant problems in molecular biology. It is common opinion that variations in the protein folding kinetics may lead to several pathologies such as prion and Alzheimer diseases. Recently, many experimental and theoretical studies have been carried out in order to describe the folding mechanism of some important proteins (Jackson, 1998; Plaxco *et al*., 1998; Fersht, 2000; Gianni *et al*., 2003; Compiani *et al*., 2004). Two main aspects of the folding process concern the kinetic order and the rate constant. The kinetic order of the protein folding indicates whether the sequence reaches its native structure through intermediate states or not. The folding rate is inversely proportional to the time that the protein needs to collapse into its 3D structure. In the last years several approaches have been implemented to estimate the logarithm of the folding rate starting from the structural information. All the methods developed so far are based on the correlation between the logarithm of the folding rate and structural parameters such as the contact order, the total contact distance or the chain length (Plaxco, 2000; Zhou and Zhou, 2002; Ivankov *et al*., 2003; Galzhiskaya *et al*., 2003; Ivankov and Finkelstein, 2004; Gong *et al*., 2003; Punta and Rost, 2005; Gromiha *et al*., 2006). These algorithms show a large value of correlation coefficient between the folding rate and different structural features; however they rarely generalize to the point of discriminating among two and multistate kinetics. It is also known that the kinetic order of the folding mechanism can be related to the folding rate (Galzhiskaya *et al*., 2003).

*To whom correspondence should be addressed.
†Present address: Structural Genomics Unit, Department of Bioinformatic (CIPF), Autopista del Saler 16, 46013 Valencia, Spain.

2 K-FOLD DESCRIPTION

K-Fold was trained to accomplish two different tasks: (1) prediction of the kinetic order of the folding process (a classification task); and (2) prediction of the $\log_{10}(k_f)$ value of the folding process (a function approximation task).

For each task, K-Fold is based on support vector machines (SVM) based on linear kernel functions.

For the classification task and for assigning the $\log(k_f)$ values we basically adopt a similar input code by identifying two labels: one represents the protein that folds without intermediate states (two-state kinetic, label is TS), the other with one or more intermediates states (multi-state kinetic, label is MS). The input vector consists of two values. The first input value accounts for the natural logarithm of the chain length (number of residues) and the second for the protein relative contact order.

K-Fold was trained and tested with a cross-validation procedure on a data set of 63 proteins, 38 of which are endowed with a TS folding mechanism. The other 25 proteins have a MS folding mechanism. Our set is essentially that previously described (Ivankov and Finkelstein, 2004) and contains 58 protein structural families distributed in the four predominant structural classes according to the SCOP classification (scop.mrc-lmb.cam.ac.uk) (a: 20.7%; b: 36.2%; c: 13.8% and d: 29.3%). The final sets are available at http://gpcr.biocomp.unibo.it/~emidio/K-Fold/dbFold.html. In order to minimize similarity among training and testing set, our cross validation is performed by dividing into five subsets, putting in the same subset proteins with related sequences, as obtained by means of the *blastclust* program.

The accuracy obtained when K-Fold is adopted as a classifier, and discriminates whether a given protein folds without (TS) or with (MS) intermediate states is 0.81 with a Matthews correlation coefficient of 0.60.

For the sake of comparison with other methods (Ivankov *et al*., 2003; Galzhiskaya *et al*., 2003) it is worth noticing that when a chain length based regression method is adopted the overall

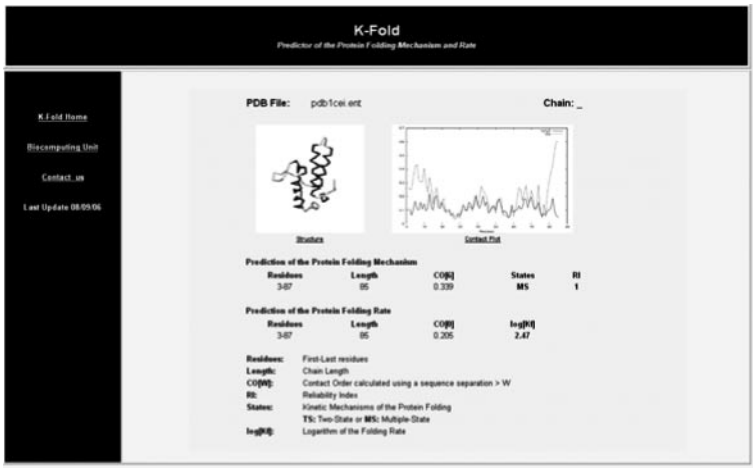

Fig. 1. Snapshot of the K-Fold output page for the prediction of the kinetic order and the logarithm of the rate constant of the folding process starting from the protein structure through its PDB code. The top side of the figure reports the picture of the protein structure (Structure, as represented with RasMol available at http://www.openrasmol.org/) and the graph representing the contact frequency and the contact order per residue (Contact Plot). The predictions of the kinetic order and of the folding rate are listed below the columns 'States' and 'log[k_f]', respectively.

discriminative accuracy on our set decreases by 7% points and concomitantly the correlation coefficient is 0.14 less. Other structural inputs were also tested by us (Capriotti and Casadio, 2006) without achieving the present performance.

K-Fold was trained/tested to predict the value of the logarithm of the folding rate, starting from the protein structure. In this case, the accuracy was evaluated by measuring the correlation between the predicted (adopting a cross validation procedure) and the observed $\log_{10}(k_f)$ values. The correlation of the predicted and experimental data is 0.74, with a SE of 1.2 (for more details see Capriotti and Casadio, 2006).

3 K-FOLD OUTPUT

Depending on the selected mode, three different outputs can be retrieved. For a specific protein structure either the kinetic order or the logarithm of the folding rate or both can be predicted. In the first two cases a table with one row is returned; for the last case the output is obtained joining the two previous outputs.

The different possible predictive options correspond to a different number of columns returned in the output table. Routinely the output page contains four columns listing respectively: the position in the sequence of the limiting residues of the protein under consideration, the chain length, the relative contact order calculated considering two different values of sequence separation and the predicted logarithm of the folding rate value ($\log_{10}(k_f)$) or the kinetic order of the folding process (TS/MS). When the kinetic order of the reaction is predicted, one more column is present in the output table that lists the reliability index value of the prediction.

When running the prediction a picture of the protein structure is also provided. In order to analyze the contribution of each residue to the global contact order the web tool reports also a graph showing the contact order and the contacts frequency per residue (Fig. 1).

ACKNOWLEDGEMENTS

This work was supported by the following grants: PNR 2001 and PNR 2003 projects (FIRB art.8) on Bioinformatics for Genomics and Proteomics and LIBI-Laboratorio Internazionale di BioInformatica, both delivered to R.C. E.C. was supported by a grant of the European Union's VI Framework Programme to the Bologna Node of the Biosapiens Network of Excellence project (contract number LSHG-CT-2003-503265).

Conflict of Interest: none declared.

REFERENCES

Jackson,S.E. (1998) How do small single-domain proteins fold? *Fold. Des.*, **3**, R81–R91.

Plaxco,K.W. *et al.* (1998) Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.*, **227**, 985–994.

Fersht,A.R. (2000) Transition-state structure as a unifying basis in protein-folding mechanism: contact order, chain topology, stability, and the extended nucleus mechanism. *Proc. Natl Acad. Sci. USA*, **97**, 1525–1529.

Gianni,S. *et al.* (2003) Unifying features in protein-folding mechanisms. *Proc. Natl Acad. Sci. USA*, **100**, 13286–13291.

Compiani,M. *et al.* (2004) Dynamics of the minimally frustrated helices determine the hierarchical folding of small helical proteins. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, **69**, 051905–051909.

Plaxco,K.W. *et al.* (2000) Topology, stability, sequence,and length: defining the determinants of two-state protein folding kinetics. *Biochemistry*, **39**, 11177–11183.

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

Ivankov,D.N. *et al.* (2003) Contact order revisited: influence of protein size on the folding rate. *Protein Sci.*, **12**, 2057–2062.

Galzitskaya,O.V. *et al.* (2003) Chain length is the main determinant of the folding rate for proteins with three-state folding kinetics. *Proteins*, **51**, 162–166.

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

Gong,H. *et al.* (2003) Local secondary structure content predicts folding rates for simple, two state proeins. *J. Mol. Biol.*, **327**, 1149–1154.

Punta,M. and Rost,B. (2005) Protein folding rates estimated from contact predictions. *J. Mol. Biol.*, **348**, 507–512.

Gromiha,M.M. *et al.* (2006) FOLD-RATE: prediction of protein folding rates from amino acid sequence. *Nucleic Acids Res.*, **34**, W70–W74.

Capriotti,E. and Casadio,R. (2006) The evalutation of protein folding rate constantis improved by predicting the folding kinetic order with a SVM based method. WSEAS. *Trans. Biol. Biomed.*, **3**, 304–310.