A machine learning approach for predicting kinetic order and rate constant of protein folding

EMIDIO CAPRIOTTI, RITA CASADIO
Biocomputing Group, Department of Biology and CIRB
University of Bologna
via Irnerio , 40126 Bologna
ITALY

Abstract: - Understanding the basic rules of protein folding is one of the most important challenges of molecular biology. In the last years several experiments have been carried out in order to study the pathway and stability of protein folding. Empirical models are available for predicting the protein folding rates, based on the linear correlation between structural protein features and folding kinetics. However no direct statistical evaluation of their prediction performance is available. Recently, a significant number of kinetic data on protein folding was published. This allows the application of machine learning methods for predicting the kinetic order and rate of protein folding starting from structural information.

In this paper we describe a support vector machine-based method suited to predict whether a protein is endowed with intermediates in the folding process and also the protein folding rate constants. Using a dataset consisting of 63 experimental protein folding data, our predictor correctly classify 78% of the folding pathways in the database and supplies an estimation of the logarithm of the folding rate constant with a correlation coefficient of 0.65. The method overcomes previous methods in optimizing the solution of folding-rate predictions. Furthermore, by predicting the presence of putative folding intermediates, it provides also a scheme for highlighting putative protein folding-mechanisms.

Key-Words: - folding kinetics, machine learning, support vector machine, contact order.

1 Introduction

In the last years, many theoretical and experimental studies have focused on the problem of describing the mechanism of protein folding [1-7]. An important result was the development of empirical models that estimate protein folding kinetics and rates. The number of proteins under investigation is rapidly increasing, allowing more data to be collected. Many proteins fold by a simple two-state transition mechanism (TS), lacking observable folding intermediates under any experimental condition. In turn, other proteins are endowed with intermediates during the folding process; their folding process is therefore classified as a multistate one (MS).

Experimental and theoretical work focused particularly on small two-state folding (TS) proteins. It was demonstrated that the logarithm of the in-water folding rates of these proteins correlates with some topological parameter as computed from their 3D structure or from that of closely related proteins, such as single point mutants or homologs with high level of sequence identity [1,8,9]. Other methods predict protein folding rates starting from the Einstein diffusion equation [10] or from the secondary structure of the protein [11]. More recent work demonstrated that the chain length is one of the main determinants of the folding rate for proteins with a multistate folding (MS) kinetics [12,13]. As a general observation, it appears that the logarithm of the folding rate correlates with structural topological parameters in TS proteins and with chain length in MS proteins.

In this paper we adopt a different perspective: we use the experimental data so far collected and, based on these observations, we develop a method to predict salient aspects of protein folding that can be directly computed starting from the protein structure.

2 Problem Formulation

The problem here addressed concerns the kinetics and mechanism of the protein folding: starting from few simple parameters derived from the protein structure, the aim is to predict important features of the folding mechanism. In particular we implement a support vector machine (SVM)-based method trained over a set of 63 proteins known with atomic resolution and whose folding pathway has been experimentally characterized to predict the logarithm of the folding rate and whether the protein folds through intermediate states or not.

2.1 Database and Tools

Our data set is derived from the supplementary material of [13]. It contains folding data determined for 63 proteins, 38 of which are endowed with a TS folding mechanism. The other 25 proteins have a MS folding mechanism. The set comprises only single-domain proteins having no S-S bonds and/or no covalently bound ligands. Furthermore the in-water folding rates (k_f) and native structure of these proteins have been established experimentally. The protein structures are available at the Protein Data Bank (www.pdb.org) [14].

The method proposed here predicts some features of the protein folding process using a SVM approach. In particular we choose the LIBSVM tools available online at the web site http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/.

The protein secondary structure was calculated with the DSSP program (http://www.cmbi.kun.nl/gv/dssp/ [15]).

Sequence clustering was performed by means of the *blastclust* program available within the BLAST suite at http://www.ncbi.nlm.nih.gov/ [16].

2.2 Protein structural parameters

In order to investigate the relationships between the folding rate constant and the protein native conformation we evaluate four structure-based parameters. The first parameter is the effective length of the protein chain (L_{eff}) defined as

$$L_{eff} = L - L_H + 3 * N_H \qquad (1)$$

where L is the chain length, L_H is the number of residues in helical conformation and N_H is the number of helices. The others topological parameters are: the contact order (CO),

$$CO = \frac{1}{N_c} \sum_{k=1}^{N_c} \Delta L_{ij} \qquad (2)$$

the relative contact order

$$RCO = \frac{1}{N \ N_c} \sum_{k=1}^{N_c} \Delta L_{ij} \qquad (3)$$

and the total contact distance

$$TCD = \frac{1}{N^2} \sum_{k=1}^{N_c} \Delta L_{ij} \qquad (4)$$

where N is the number of amino acid residues of a protein, N_c is defined as total number of contacts and $\Delta L_{ij} = |i-j|$.
The number of contacts is evaluated considering all the residues that have two heavy atoms within a given value of cut-off radius R and at a given sequence separation (w).

2.3 The predictor

The method addresses two different tasks: (1) the prediction of the existence of intermediate states in protein folding and (2) the prediction of the logarithm of the folding rate value. The former case is a classification task, discriminating whether for a given protein the folding pathway is or is not endowed with intermediate states; the latter one in turn is a fitting-regression task for estimating the folding rate. To address the two tasks, we developed a method based on support vector machines and relying on the same input for testing different kernel functions. Also, different SVMs explore different protein features. SVMs take two inputs for a given protein: the chain length and, one at a time, the four structured-based parameters described above (Eqn 1-4). We found that the best performing predictor was the one having as input the protein chain length and the contact order, tested by splitting the dataset in five parts and adopting a cross-validation procedure. The methods were then optimized trying different values of cut-off radius (R) and of sequence separation (w).

2.4 Scoring the classification performance

All the results obtained with our systems are evaluated using a cross-validation procedure on the data pertaining to the 63 proteins. The dataset was divided in 5 subsets, putting in the same set proteins with the same PDB code and proteins with related sequences as obtained by means of the *blastclust* program by adopting the default value of length coverage equal to 0.9 and the score coverage threshold equal to 1.75.

The efficiency of the predictor is scored using the statistical indexes defined in the following. The overall accuracy is:

$$Q2 = \frac{p}{N} \qquad (5)$$

where p is the total number of correctly predicted folding mechanisms and N is the total number of proteins.
The Matthews correlation coefficient MC is defined as:

$$MC(s) = \frac{\mathrm{p(s)n(s) - u(s)o(s)}}{\mathrm{D}} \qquad (6)$$

where D is the normalization factor $[[p(s)+u(s)] [p(s)+o(s)] [n(s)+u(s)] [n(s)+o(s)]]^{1/2}$, for each class s (TS and MS, for two-state and multistate folding processes, respectively); p(s) and n(s) are the total number of correct predictions and correctly rejected assignments, respectively, and u(s) and o(s) are the numbers of under and over predictions.

Finally, it is very important to assign a reliability score to each SVM prediction. Using one SVM output this is obtained by computing:

$$\mathrm{Re}\,l(i) = 20 * abs[O(i) - 0.5] \qquad (9)$$

2.5 Scoring the regression performance

The quality of the prediction when evaluating the protein folding constant rates was assessed by computing the Pearson linear correlation coefficient r and the associated value of the standard error σ.

3 Problem Solution

In order to solve the tasks discussed in section 2.3 we developed different support vector machines. Taking advantage of previous studies, each of the SVMs considers two important protein features: (1) sequence length and (2) the four structural parameters described above. The best performing predictor was then optimized testing different values of cut-off radius (R), different sequence separation values (w) and different kernel functions. We found that the best performance was achieved by a SVM endowed with a linear kernel function $K(x_i,x_j)=x_iTx_j$ (data not shown).

3.1 Structural parameter optimization

Previous studies have highlighted in proteins the correlation between folding kinetics and structural parameters as described in section 2.2 [6,8,9,13]. Table 1 lists the scoring performance of each method when predicting the logarithm of the folding rate and the folding kinetics.

		L_{eff}	CO	RCO	TCD
Prediction of Folding States	MC	0.15	0.42	0.27	0.36
	Q2	57.1	73.2	65.9	69.8
Prediction of $\log(k_f)$	r	0.45	0.64	0.45	0.63
	σ	1.57	1.39	1.57	1.37

Table 1. Scoring the SVM method. The first two rows list the accuracy (Q2) and the Matthew's correlation coefficient (MC) of the four methods that include in the SVM input one of the different structural parameters and the sequence length. The four SVM labeled with the name of the relative structural parameter, are tested in the binary classification between of two-state and multistate folding mechanism. In the last two rows the correlation coefficient (r) and the standard error (σ) of the previous methods in the prediction of the logarithm of the folding rate (k_f) are reported.

3.2 Optimization of the cut-off radius

The results shown in Table 1 indicate that the best SVM method has as input the sequence length and the contact order (see column CO). For this method we tested different values of the cut-off radius. In Table 2, the scoring indexes for the two previous tasks are shown as a function of the radius value ranging from 4 to 12 Å.

		4	6	9	12
Prediction of Folding States	MC	0.42	0.31	0.48	0.40
	Q2	73.2	67.1	75.6	72.0
Prediction of $\log(k_f)$	r	0.64	0.61	0.65	0.64
	σ	1.39	1.44	1.35	1.38

Table 2. **Scoring SVMs as a function of the cut-off radius**. Here the method takes as input protein sequence length and contact order. The last structural parameter is calculated using a cut-off radius ranging from 4 to 12 Å . The first two rows list the quality of the prediction in the classification task; the last two rows show the quality of the prediction of the logarithm of the folding rate (k_f).

3.3 Sequence separation optimization

When considering the protein folding mechanism, an important issue is the different contribution of local and non-local interactions. It is well known that local interactions involved in the formation of particular motifs of secondary structure are established between residues with a sequence separation below 4 residues that is about the residue distance of one turn of an α-helix structure. Therefore increasing the value of w in the calculation of CO, we go beyond local interactions and include also contacts between residues that may contributes to non-local interactions during the folding process. We address this task by evaluating the contact order as a function of sequence separation; the best performing implementation of SVMs was consequently optimized and the results are shown in Table 3

		0	2	4	6	8
Prediction of Folding States	MC	0.48	0.5	0.46	0.53	0.42
	Q2	75.6	76.1	74.6	77.7	73.1
Prediction of $\log(k_f)$	r	0.65	0.6	0.61	0.58	0.6
	σ	1.35	1.52	1.29	1.45	1.41

Table 3. Local vs global interactions. In this table we report the accuracy of the best methods (cut-off radii 9 Å) for different values of a sequence separation (w), spanning from 0 to 8 residue, when evaluating the contact order number. In other words, we consider only contact between residue i and j if $|i-j|>=w$. The efficiency of the predictions for the two tasks are scored using the same measures reported in Table 1.

3.4 Prediction of the folding mechanism.

From our results we conclude that the best method for the binary classification between the two-state and the multistate folding mechanism takes as input the sequence length and the contact order. The best discrimination between TS and MS proteins is obtained when the contact order value is calculated considering a cut-off radius of 9 Å and a sequence separation ≥6 residues. In figure 1 we report the accuracy (Q2) and the Matthew's correlation coefficient (MC) as a function of the reliability index (RI).

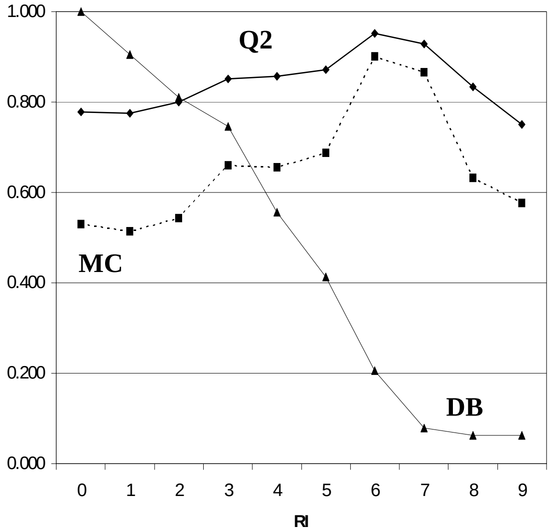

Figure 1. Accuracy (Q2) and Matthew's correlation coefficient (MC) as o function of the reliability index (RI). DB is the fraction of the dataset with a reliability value higher or equal to a given threshold.

3.5 Prediction of the logarithm of the folding rate

Similar to the classification task, the regression task for the prediction of the logarithm of the folding rate is optimized considering as input the sequence length and the contact order. The results of our method are shown in figure 2

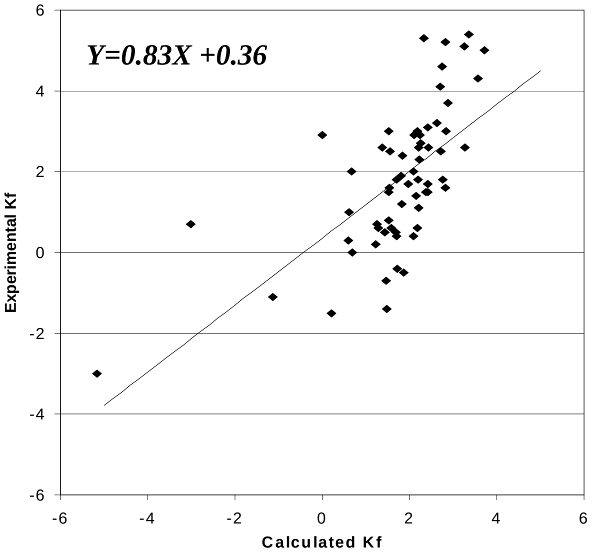

Figure 2. **Value of the logarithm of the folding rate (k_f) versus its experimental value**. The correlation coefficient for the best method previously described is 0.65 and the standard error is 1.35. We also reported the equation of the linear best fit.

4 Conclusion

This work represents a first attempt to address the problem of the prediction of the folding mechanism using a machine learning approach. In particular we try to predict whether the folding process follows a two-state or a multistate mechanism and the logarithm of the folding rate considering only few simple inputs: the length of the protein sequence and the contact order, as calculated according to the eq. (2). This is the first time, at the best of our knowledge, that a statistical evaluation of the problem is provided. We optimize our method considering different values of the cut-off radius and introducing a sequence separation for the calculation of the contact order (CO) from the protein structure, in order to discriminate local versus non local interactions. Our approach allows to generalize on the given examples since it is tested adopting a cross-validation procedure. We find that the best predictive performance is achieved when the value of the contact order is calculated using a cut-off radii of 9 Å and a sequence separation larger or equal to 6, suggesting that non local more than local interactions are important in determining the parameters at hand for the given protein set.

With our method the prediction of possible intermediate states during the folding process reaches accuracy of 78% with a significant Matthew's correlation coefficient of 0.53. Furthermore, when predictions with a reliability index value ≥3 are considered, the SVM method increases its accuracy to 85% and its correlation to 0.66 over 75% of the database. Results in Tab. 2 indicate that for discriminating between TS and MS folding mechanisms, contacts between residues with sequence separation ≥6 are important. In turn, for predicting the value of the logarithm of the folding rate the highest score is obtained considering all the contacts. On the contrary, with respect to the classification between TS and MS proteins, the regression task for the prediction of the logarithm of k_f, performs better when local and non local interactions are considered taking also into account contacts with sequence separation less or equal then 6. In this particular task our best method reaches a significant correlation coefficient of 0.65 with a related standard error of 1.35. These values can be considered satisfactory, since they are obtained with only two element vectors as input in the training of the SVM and since the method is tested using a cross-validation procedure.

This work is a good starting point for building more accurate predictors of the folding mechanism considering a larger number of features in the training of the SVM and merging the two methods here developed that are related to different aspects of the protein folding process.

References:

[1] Jackson SE. How do small single-domain proteins fold? *Fold Des*, Vol. 3, 1998, pp. R81-R91.

[2] Plaxco KW, Simons KT, Baker D. Contact order, transition state placement and the refolding rates of single domain proteins. *J Mol Biol*, Vol. 227, 1998, pp. 985 –994.

[3] Fersht. AR 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*, Vol. 97, 2000, pp. 1525-1529.

[4] Gianni S, Guydosh NR, Khan F, Caldas TD, Mayor U, White GWN, DeMarco ML, Daggett V, Fersht AR. Unifying features in protein-folding mechanisms. *Proc. Natl. Acad. Sci. USA*, Vol. 100, 2003, pp. 13286-13291.

[5] Compiani M, Capriotti E, Casadio R. Dynamics of the minimally frustrated helices determine the hierarchical foldingof small helical proteins. *Phys Rev E Stat Nonlin Soft Matter Phys*, Vol. 69, 2004, pp. 051905-051909.

[6] Plaxco KW, Simons KT, Ruczinski I, Baker D. Topology, stability, sequence,and length:defining the determinants of two-state protein folding kinetics. *Biochemistry* Vol. 39, 2000, pp. 11177 –11183.

[7] Garbuzynskiy SO, Finkelstein AV, Galzitskaya OV. Outlining folding nuclei in globular proteins. *J Mol Biol*, Vol. 336, 2004, pp. 509-525.

[8] Zhou H, Zhou Y. Folding rate prediction using total contact distance. *Biophys J*, Vol. 82, 2002, pp. 458-463.

[9] Ivankov DN, Garbuzynskiy SO, Alm E, Plaxco KW, Baker D, Finkelstein AV. Contact order revisited: influence of protein size on the folding rate. *Protein Sci.* Vol. 12, 2003, pp. 2057-2062.

[10] Debe DA, Goddard WA 3rd. First principles prediction of protein folding rates. *J Mol Biol*, 1999, pp. 619-625.

[11] Gong H, Isom DG, Srinivasan R, Rose GD. Local secondary structure content predicts folding rates for simple, two-state proteins. *J Mol Biol*, Vol. 327, 2003, pp. 1149-1154.

[12] Galzitskaya OV, Garbuzynskiy SO, Ivankov DN, Finkelstein AV. Chain length is the main determinant of the folding rate for proteins with three-state folding kinetics. *Proteins*, Vol. 51, 2003, pp. 162-166.

[13] Ivankov DN, Finkelstein AV. Prediction of protein folding rates from the amino acid sequence predicted secondary structure. *Proc Natl Acad Sci USA*, Vol. 101, 2004, pp. 8942-8944.

[14] Berman, H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N., and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res*, Vol. 28, pp. 235-242.

[15] Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern of hydrogen-bonded and geometrical features. *Biopolymers*, Vol. 22, pp. 2577-2637.

[16] Altschul SF, Madden TL, Shaffer AA, Zhang J, Zhang Z, Miller W, Lipman DI, Gapped BLAST and PSI-LAST a new generation of database search programs. *Nucleic Acids Res*, Vol. 25, 1997, pp. 3389-3402.