Amino Acids (2012) 42:271–283
DOI 10.1007/s00726-010-0805-y

ORIGINAL ARTICLE

Determination of protein folding kinetic types using sequence and predicted secondary structure and solvent accessibility

Hua Zhang · Tuo Zhang · Jianzhao Gao · Jishou Ruan · Shiyi Shen · Lukasz Kurgan

Received: 17 February 2010 / Accepted: 1 November 2010 / Published online: 17 November 2010

Abstract Proteins fold through a two-state (TS), with no visible intermediates, or a multi-state (MS), via at least one intermediate, process. We analyze sequence-derived factors that determine folding types by introducing a novel sequence-based folding type predictor called FOKIT. This method implements a logistic regression model with six input features which hybridize information concerning amino acid composition and predicted secondary structure and solvent accessibility. FOKIT provides predictions with average Matthews correlation coefficient (MCC) between 0.58 and 0.91 measured using out-of-sample tests on four benchmark datasets. These results are shown to be competitive or better than results of four modern predictors. We also show that FOKIT outperforms these methods when predicting chains that share low similarity with the chains used to build the model, which is an important advantage given the limited number of annotated chains. We demonstrate that inclusion of solvent accessibility helps in discrimination of the folding kinetic types and that three of the features constitute statistically significant markers that differentiate TS and MS folders. We found that the increased content of exposed Trp and buried Leu are indicative of the MS folding, which implies that the exposure/burial of certain hydrophobic residues may play important role in the formation of the folding intermediates. Our conclusions are supported by two case studies.

Keywords Folding · Folding kinetic types · Folding rate · Solvent accessibility · Secondary structure

H. Zhang
School of Computer Science and Information Engineering, Zhejiang Gongshang University, Hangzhou, People's Republic of China

H. Zhang · T. Zhang · J. Gao · J. Ruan · S. Shen
College of Mathematical Science and LPMC, Nankai University, Tianjin, People's Republic of China

H. Zhang · T. Zhang · L. Kurgan (✉)
Department of Electrical and Computer Engineering, University of Alberta, ECERF (9107 116 Street), Edmonton, AB T6G 2V4, Canada
e-mail: lkurgan@ece.ualberta.ca

T. Zhang
Indiana University School of Informatics, Indiana University-Purdue University, Indianapolis, IN, USA

J. Ruan · S. Shen
Chern Institute of Mathematics, Tianjin, People's Republic of China

Introduction

Protein folding, which spans processes between an initial random coil conformation and the functional native structure, occurs through a diverse range of pathways that may include intermediate states (Anfinsen 1973; Udgaonkar 2008). Characterization and analysis of these complex processes pose substantial challenges for both experimental and computational methods (Dill et al. 2008). The applicable experimental techniques include optical spectroscopies and laser-induced temperature-jump (Callender et al. 1998; Schuler et al. 2002), Ψ and Φ values analyses (Sosnick et al. 2004), hydrogen exchange (Krishna et al. 2004; Maity et al. 2005), and NMR relaxation (Lindorff-Larsen et al. 2005). In spite of the availability of the wide array of experimental methods the kinetic data are accumulated at a low rate. The two main repositories, KineticDB (Bogatyreva et al. 2009) and Protein Folding Database (PFD) (Fulton et al. 2007), include data on only about 90 proteins,

which is due to the difficulty in experimental determination of the protein kinetics. The folding kinetics is also studied in silico using molecular dynamics simulations, but due to high computational cost this analysis could be applied only to peptides or small proteins (Scheraga et al. 2007). Another feasible alternative is to use the available experimental data to build computational models that can be utilized to provide insights into certain folding characteristics and to predict them for the unsolved protein chains. A prime example of such characteristic is folding rate. Several works have investigated the relation of the folding rate with structural characteristics of the native fold including contact order (Plaxco et al. 1998), long-range order (Gromiha and Selvaraj 2001), absolute contact order (Ivankov et al. 2003), relative contact order (Capriotti and Casadio 2007), geometric contacts (Ouyang and Liang 2008), multiple contact index (Gromiha 2009), and structural compactness (Galzitskaya et al. 2008; Ivankov et al. 2009). Other factors, such as chain length (Galzitskaya et al. 2003), secondary structure (Gong et al. 2003; Huang et al. 2007), effective length of predicted secondary structure (Ivankov and Finkelstein 2004), predicted contact maps (Punta and Rost 2005), amino acid (AA) indices (Huang and Gromiha 2008), AA composition (Ma et al. 2006) and a combination of predicted secondary structure, protein chain, and physiochemical properties of residues (Jiang et al. 2009; Shen et al. 2009), have been also used to analyze and predict folding rates. Another aspect which is related to the rate (Galzitskaya et al. 2003) is the type of folding process, which includes two-state (TS) and multi-state (MS) folding (Finkelshtein and Galzitskaya 2004; Kamagata et al. 2004). The TS folding is a reversible process that has no visible intermediates (Jackson 1998; Finkelshtein and Galzitskaya 2004), while MS proteins fold via at least one or more intermediates where the folding process follows a stepwise assembly procedure (Jackson 1998; Maity et al. 2005; Feng et al. 2005). The knowledge of the folding kinetic types was shown to improve the quality of the prediction of the folding rates (Huang and Cheng 2008) and finds applications in determination of the folding intermediates (Ma et al. 2007). However, only a few works focused on the characterization and prediction of the folding kinetic types. Capriotti and Casadio (2007) used chain length and relative contact order computed from the native fold to predict the folding kinetic types. Ma et al. (2007) performed a comparative investigation into the relationship of the kinetic types and AA composition of the protein chain and the topology of the native fold. Two sequence-based prediction models, one based on physicochemical residue properties (Huang and Gromiha 2008) and another based on chain length (Huang and Cheng 2008), were proposed in 2008. Most recently, compactness of the native fold was demonstrated to explain some differences in the folding mechanisms (Galzitskaya et al. 2008).

Although the folding rates were shown to depend on the secondary structure (Gong et al. 2003; Ivankov and Finkelstein 2004; Huang et al. 2007) and although recent work suggests that folding kinetic types depend on the characteristics of the protein surface (Galzitskaya et al. 2008), these two factors were not utilized to predict/characterize folding kinetic types. There is also evidence showing that specific residue mutations can result in switching between the TS and MS processes (Jackson 1998; Inaba et al. 2000; Viguera and Serrano 2003; Cranz-Mileva et al. 2005). The Φ value analysis of chymotrypsin inhibitor 2 revealed a relation between folding rate and stability of the native fold (Fersht 2000). In another study, a single mutation in hen egg-white lysozyme resulted in a less stable structure than that of the wild-type (Zhou et al. 2007). Finally, recent studies show that surface Trp residues strongly contribute to the folding stability (Klein-Seetharaman et al. 2002; Zhou et al. 2007; Zhang et al. 2009), and thus they may have impact on the folding kinetics. At the same time, factors related to point mutations, secondary structure, and solvent accessibility are not utilized by the existing sequence-based predictors of the folding kinetic types (Huang and Gromiha 2008; Huang and Cheng 2008; Ma et al. 2007), which apply only information concerning the length and composition of the protein chain. To this end, we propose a novel sequence-based folding type predictor named FOKIT (*fo*lding *ki*netic *t*ype). FOKIT hybridizes information concerning chain length and composition, solvent accessibility predicted with Real-SPINE (Dor and Zhou 2007), and secondary structure predicted with PSI-PRED (Bryson et al. 2005), to compute six features that are fed into a logistic regression classifier. Empirical tests demonstrate that the proposed method outperforms existing solutions. The FOKIT's prediction model provides interesting insights into the folding kinetics, which are demonstrated using case studies that draw from existing experimental findings.

Materials and methods

Datasets

A dataset composed of 85 proteins, named H85, which was recently introduced in by Huang and Cheng (2008), is used to design the prediction model. The H85 dataset includes 60 TS and 25 MS proteins. We also utilize three benchmark datasets, M85, C63, and G77, to compare with the existing predictors, see Table 1. The four datasets share a significant portion of the chains since they were created mostly using entries from the KineticDB (Bogatyreva et al.

Table 1 Datasets used to develop and evaluate the proposed predictor

Reference	Abbreviated dataset name	# TS proteins	# MS proteins
Huang and Cheng (2008)	H85	60	25
Ma et al. (2007)	M85	43	42
Capriotti and Casadio (2007)	C63	38	25
Huang and Gromiha (2008)	G77	51	26
This work	S17	7	10

2009) and PFD (Fulton et al. 2007) databases, which when combined include only around 90 proteins.

We additionally tested our predictor on chains with low similarity to the chains in the H85 dataset that were used to develop the prediction model. Such test was not attempted in the past, i.e., each of the above four datasets which were used to evaluate the existing methods includes similar chains. To reduce the sequence identity, BLASTCLUST (Altschul et al. 1997) was applied to the union of the four datasets with the local identity threshold of 25% (-S 25). The new dataset was constructed by selecting one chain from each of the clusters that contained no sequences from the H85 dataset. The resulting set, called S17, includes seven TS and ten MS chains that have up to 25% local identity with each other and also with the chains in the H85 dataset. This dataset is comprised of 11, 2, and 4 proteins from the M85, C63, and G77 datasets, respectively, and represents virtually all available chains that have the kinetic annotations and that are dissimilar to the chains in the H85 dataset.

The five datasets, H85, M85, C63, G77, and S17 are available for download at http://biomine.ece.ualberta.ca/FOKIT/FOKIT.htm.

Logistic regression

Logistic regression (LogR) is a method suitable to model a logistic relationship between the discrete outcome (in our case the binary annotations of the folding kinetic types) and the numeric feature vector. This method has been applied to various problems in computational biology, such as classification of antibacterial activity (Cronin et al. 2002), prediction of flexible regions in proteins (Chen et al. 2007), and prediction of the folding kinetic types (Huang and Cheng 2008). The binary logistic regression model has the following form:

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k)}},$$

where $X = (X_1, X_2, \ldots, X_k)$ is the feature vector, β_0 is the regression constant, $\beta_1, \ldots, \beta_k$ are the regression coefficients for of $X_1, \ldots, X_k$, respectively. The outcomes $p > 0.5$ and $p < 0.5$ indicate that the vector X is categorized as positive (TS protein) and negative (MS protein), respectively.

Performance evaluation

Prediction performance is assessed using four quality indices including sensitivity (the ratio between the number of correct predictions for TS proteins and the total number of the actual TS proteins), specificity (the ratio between the number of correct predictions for MS proteins and the total number of the actual MS proteins), the overall accuracy, and Matthews correlation coefficient (MCC) (Matthews 1975):

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100\%$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \times 100\%$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \times 100\%$$

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FN})(\text{TP} + \text{FP})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}$$

where true positives (TP) and true negatives (TN) correspond to correctly predicted TS and MS proteins, respectively, false positives (FP) denote MS proteins predicted as TS proteins, and false negatives (FN) denote TS proteins predicted as MS. The MCC measure ranges between -1 and 1, where -1 corresponds to all incorrect predictions, 0 to random predictions, and 1 to all correct predictions. This measure accommodates for the unbalanced number of TS and MS proteins, i.e., if all chain in the H85 dataset would be classified as TS then the accuracy would equal $100 \times 60/85 = 70.6\%$ (instead of "expected" 50%), while the MCC would be 0.

The performance is tested using n fold cross validation (nCV) tests with multiple runs (to improve validity of the results considering small sizes of the datasets) on the H85, M85, C63, and G77 datasets. In the nCV, chains are randomly divided into n subsets with the same numbers of sequences, and the test is repeated n times, each time using one subset to test the prediction model and the remaining $n - 1$ subsets to establish the model. Execution of one nCV is called a run and the n subsets for the run are named a seed. We performed fivefold cross validation (5CV) following the work by Capriotti and Casadio (2007) and by Huang and Cheng (2008), but we executed ten runs using ten different randomly created seeds. The sensitivity, specificity, accuracy, and MCC are computed for each run and then averaged over the ten runs. The jackknife test (JKT), also called the leave-one-out test, assumes that n is the total number of sequences in the dataset. We execute only one run since each run would give the same result.

Table 2 Summary of the considered features, where $y = \{C, H, E\}$ denotes the three secondary structure states, $x = \{A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V\}$ denotes the 20 AA types, and $h = \{0.1, 0.2, 0.3, 0.4, 0.5\}$ denotes the cutoff used to categorize the buried/exposed residues based on their relative solvent accessibility

Category	Feature description	Abbreviation	No. of features
Sequence based	Length	L	1
Secondary structure based	Effective length	Eff_L	1
	Content of secondary structure y	$SSCon_y$	3
	Segment content of secondary structure y	$SegCon_y$	3
Average RSA based	Average RSA of the residues with AA type x	$AveRSA\_AA_x$	20
	Average RSA of the residues with secondary structure type y	$AveRSA\_SS_y$	3
	Average RSA of the residues with AA type x and secondary structure type y	$AveRSA\_AA_x\_SS_y$	60
Amino acid composition based	Composition of AA x	AAC_x	20
	Composition of AA x with secondary structure type y	$AAC_x\_SS_y$	60
	Composition of AA x with RSA value $\geq h$ (i.e., the residue is assumed exposed with respect to the threshold h)	$AAC_x\_Ex_h$	100
	Composition of AA x with RSA value $< h$ (i.e., the residue is assumed buried with respect to the threshold h)	$AAC_x\_Bu_h$	100

The results reported for the existing sequence-based folding type predictors are based either on the 5CV or JKT. On the contrary, we report results based on both test procedures for all datasets. In addition, Ma et al. (2007) introduced a variation of the Jackknife test (JKT^v), in which one sequence is randomly chosen to test the model and this is repeated 1,000 times. We also report this test, but we limit it to the M85 dataset that was used by these authors.

Features

The FOKIT method converts the input protein sequence into a set of numerical features that are fed into the logistic regression classifier to generate prediction of the folding kinetics type. The features include the chain length and the AA composition which were previously shown to discriminate between TS and MS proteins (Huang and Cheng 2008; Ma et al. 2007). We also introduced novel features that utilize the predicted secondary structure (PredSS), the predicted relative solvent accessibility (PredRSA), the combination of the above. The features are divided into four categories, sequence based, secondary structure based, average RSA based, and AA composition based (see Table 2).

The motivation for using PredSS comes from several studies which have shown that the knowledge of the secondary structure (SS) helps to predict the folding rates (Gong et al. 2004; Ivankov and Finkelstein 2004; Huang et al. 2007; Jiang et al. 2009). Although differences in the secondary structure content values between TS and MS folders were shown not to be statistically significant (Ma et al. 2007), we examined other types of features computed using information concerning secondary structure segments and combining the secondary structure and solvent accessibility/AA composition. The PSIPRED server (Bryson et al. 2005) was used to predict three-state secondary structures. This was motivated by the wide-spread usage of this predictor in related methodologies (Song and Burrage 2006; Chen and Kurgan 2007; Zhang et al. 2008) and in prediction of the folding rate (Ivankov and Finkelstein 2004; Jiang et al. 2009; Shen et al. 2009). The effective length was calculated as (Ivankov and Finkelstein 2004)

$$\text{Eff\_L} = L - L_H + 3 \times N_H$$

where L is the chain length, L_H is the number of residues in predicted helix conformation, and N_H is the number of predicted helix segments. The content of a given SS is defined as the ratio of the number of residues in this conformation to the sequence length. The segment content of a given SS type is the percentage of the number of sequence segments in this conformation to the total number of SS segments in the sequence.

The relative solvent accessibility (RSA) is defined as the solvent accessible surface area (ASA) of a given residue normalized by the ASA of this residue in an extended tripeptide, Ala-X-Ala, conformation (Ahmad et al. 2003). The ASA values were predicted using Real-SPINE program (Dor and Zhou 2007), which is motivated by the high quality of these predictions (Zhang et al. 2009). The RSA values are often used to differentiate between the interior and the surface of proteins by setting a cutoff. For a given cutoff h, the residue with RSA $\geq h$ are considered to be solvent exposed; otherwise, they are assumed to be buried. The burial/exposure of certain residue types may play

important role in the folding stability, especially for hydrophobic residues, which had been shown in several cases (Calloni et al. 2003; Dong et al. 2008; Esposito et al. 2008; Ricagno et al. 2009). We computed average RSA (AveRSA) value of the residues of certain AA type, in a given predicted secondary structure conformation, and of the residues of certain AA type in a given predicted secondary structure state.

The AA composition based features include composition of the 20 AA types in the input sequence, the composition of the residues of certain AA type in a given predicted secondary structure conformation, and the composition of the residues of certain AA type which are either buried or exposed based on different RSA cutoffs (see Table 2).

Feature selection

The considered feature set is composed of 371 descriptors. We perform feature selection since some of these features could be irrelevant to the prediction/characterization of kinetic types. We utilized two types of selection methods: filter-based approach that evaluates the features based on a strength of their relation with the annotation of the folding kinetics type, and wrapper-based approach, which selects features that provide favorable prediction quality when used with a given prediction method. We expect that the wrapper-based approach will lead to better predictions since it was previously shown to outperform the filter-based approach when selecting features for a subsequent classification (Kohavi and John 1997). A forward, best-first search over the sets of features ranked using the filter- and wrapper-based methods was performed. We use a correlation-based filter (Yu and Liu 2003) and three wrappers with logistic regression (LogR), and two support vector machines (SVM) (Vapnik 1998) with different kernel types as the base predictors. The filter and the regression models were computed using Weka platform (Witten and Frank 2005) and we used LIBSVM library (Hsu and Lin 2002) to implement the SVM models.

For the wrapper-based methods, the feature sets that lead to a higher average Matthews correlation coefficient (MCC) (see "Materials and methods" section for the definitions of quality measures) when used to predict folding kinetics type with either LogR or SVMs are selected. The computation of the MCC involves out-of-sample tests on the H85 dataset. More specifically, we execute ten random seeds of fivefold cross validation (5CV) (see "Materials and methods" section for the definitions of datasets and test procedures) and we use the average MCC to rank features. We start with one feature that gives the largest MCC and we add the second feature (among the remaining 370 features) which results in the best average MCC. This is performed incrementally until adding an additional feature does not improve the average MCC value when performing prediction with LogR and SVMs. Additionally, the SVM classifiers require parametrization of the complexity constant C and the kernel function. We consider two kernel types, radial basis function (RBF) $K(x_i, x_j) = \exp(-\gamma\|x_i - x_j\|^2)$ where γ is the width of the RBF function, and polynomial $K(x_i, x_j) = (x_i \times x_j)^d$ where d is the degree. We perform a grid search for the best parameters. For the RBF kernel, $C = \{2^{-5}, 2^{-4}, \ldots, 2^4, 2^5\}$ and $\gamma = \{2^{-5}, 2^{-4}, \ldots, 2^4, 2^5\}$, and for polynomial kernel $C = \{2^{-2}, 2^{-1}, \ldots, 2^9, 2^{10}\}$ and $d = \{1, 2, 3\}$; we adjusted the values of C for each kernel to make sure that the optimal parameters are inside the grid. The parameterization is performed again each time an additional feature is added to the set of the selected features.

In the case of the filter approach, the features characterized by high correlation with the annotation of the folding kinetics type and low intercorrelation are selected. Similarly, as for the wrapper-based approach we estimate the correlations using ten runs of 5CV. This results in total of 50 feature sets, and we select the features that appear in majority of these sets, i.e., in at least 25 out of 50 sets of selected features.

Using the RBF kernel-based SVM the following features are selected: L, $AAC_{L}\_SS_{C}$, $AveRSA\_AA_{G}\_SS_{H}$, $AAC_{L}\_Ex_{0.5}$ and AAC_{M}, and the corresponding optimal C and γ values are 16 and 2, respectively. For the polynomial kernel-based SVM the selected features include L, $AAC_{L}\_SS_{C}$, $SegCon_{C}$, $AveRSA\_SS_{H}$, $AAC_{M}\_SS_{H}$, $AveRSA\_AA_{D}\_SS_{C}$, $AAC_{H}\_SS_{H}$, $AveRSA\_AA_{L}$,$AveRSA\_AA_{C}\_SS_{C}$, $AAC_{H}\_SS_{E}$, $AAC_{A}\_SS_{H}$, $AAC_{L}\_Ex_{0.2}$, and the corresponding optimal C equals 128 and d is 1. The features chosen using the LogR include L, $AveRSA\_AA_{N}\_SS_{E}$, $AveRSA\_AA_{E}\_SS_{E}$, $AAC_{E}\_SS_{E}$, $AAC_{L}\_Bu_{0.1}$, and $AAC_{W}\_Ex_{0.2}$, and the features selected using the filter-based approach are L, Eff_L, $AveRSA\_AA_{E}$, and $AAC_{V}\_Ex_{0.2}$. We note that each selection procedure leads to a different feature set; the only intersecting feature is the chain length L.

We chose among these alternative designs by comparing the quality of prediction of the folding rate types using each of the above four feature sets. The three feature sets selected using the wrapper-based methods are used together with their corresponding predictors. We utilize the same three predictors, two SVMs and LogR, with the feature set chosen using the filter-based method. Similar to the feature selection, for the SVM-based predictors we run a grid search with $C = \{2^{-5}, 2^{-4}, \ldots, 2^4, 2^5\}$ and $\gamma = \{2^{-5}, 2^{-4}, \ldots, 2^4, 2^5\}$ for the RBF kernel and with $C = \{2^{-2}, 2^{-1}, \ldots, 2^9, 2^{10}\}$ and $d = \{1, 2, 3\}$ when using the polynomial kernel to optimize the average, over the ten random seeds of 5CV on the H85 dataset, MCC. Table 3

Table 3 The comparison of the prediction performance of the logistic regression (LogR)-based wrapper, two support vector machine (SVM)-based wrappers (with the RBF and polynomial kernels), and three predictors that use the filter-based selected features and the LogR, SVM with the RBF kernel, and SVM with the polynomial kernel predictors, respectively

Prediction method	Sensitivity	Specificity	Accuracy	MCC
LogR-based wrapper	0.975 ± 0.023	**0.912** ± 0.017	**0.956** ± 0.018	**0.895** ± 0.044
SVM-based wrapper with polynomial kernel	0.998 ± 0.005	0.820 ± 0.027	0.946 ± 0.008	0.870 ± 0.019
SVM-based wrapper with RBF kernel	**1.000** ± 0.000	0.784 ± 0.020	0.936 ± 0.006	0.848 ± 0.014
SVM with polynomial kernel using filter-based selected features	0.983 ± 0.000	0.688 ± 0.016	0.896 ± 0.005	0.746 ± 0.012
SVM with RBF kernel using filter-based selected features	0.983 ± 0.000	0.680 ± 0.000	0.894 ± 0.000	0.740 ± 0.000
LogR using filter-based selected features	0.922 ± 0.015	0.752 ± 0.016	0.872 ± 0.011	0.687 ± 0.025

The results are based on ten fivefold cross validation runs on the H85 dataset and the averages and the corresponding standard deviations are shown. The methods are sorted by the average MCC values in the descending order, and the best values for each quality index (see "Materials and methods" section for details) are given in bold

compares the results. As expected, the performance of the wrapper-based designs is better than that of the filter-based methods. The predictions with the SVM that applies the polynomial kernel and with the logistic regression have the best and comparable quality with accuracy around 0.95 and MCC close to 0.9. Both of these solutions are based on linear methods, i.e., the SVM uses a linear polynomial. We use the LogR-based design to implement the proposed folding type predictor since this method provides high-quality predictions, and it has simpler underlying prediction model and uses 50% fewer features when compared with the second-best SVM.

For the best performing wrapper-based with LogR selection we also validated the selection procedure using jackknife test instead of the 5CV tests (see Fig. 1). The same set of six features was selected using both test types. Except for the chain length, the other five features used by the LogR-based solution are novel when compared with features used by the existing predictors; they hybridize information coming from RSA, SS, and AA composition. The number of the selected features is comparable with the number used by the other methods. More specially, the most recent method uses just the chain length (Huang and Cheng 2008), the FOLD-RATE Q (Huang and Gromiha 2008) applies ten features, and K-Fold (Capriotti and Casadio 2007) uses two features. The feature values for all considered datasets are available for download at http://biomine.ece.ualberta.ca/FOKIT/FOKIT.htm.

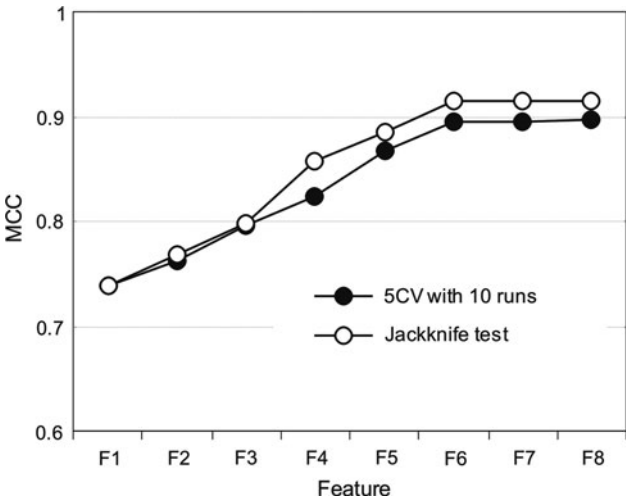

Fig. 1 The improvements of MCC values (y axis) along with the increasing number of selected features (x axis) for the performed wrapper-based feature selection. A forward, best-first search was executed using both 10 5CV runs and jackknife tests on the H85 dataset. The features F1 to F8 are L, AveRSA_AA_N_SS_E, AveRSA_AA_E_SS_E, AAC_E_SS_E, AAC_L_$Bu_{0.1}$, ACC_W_$Ex_{0.2}$, AAC_T_$Bu_{0.1}$, and AAC_T_$Bu_{0.5}$, respectively, in the order of their inclusion in the feature selection procedure

Results and discussion

Comparison with chain length-based predictions

We compare FOKIT against predictions based solely on the chain length, (see Fig. 2), which is motivated by recent results that suggest that the length is sufficient for accurate predictions (Huang and Cheng 2008). The usage of the five additional features in the proposed method is shown to result in substantial improvements. For instance, the MCC on the H85 dataset improves from 0.738 to 0.895 for the 5CV test, and from 0.738 to 0.914 for the jackknife test. We note that the chain length provides relatively accurate predictions and that the H85 dataset is the easiest and the M85 dataset is the most difficult to predict using the length. FOKIT provides improvements on all four datasets, and it follows the same trend with respect to difficulty in predicting the four datasets.

Comparison with existing methods

Table 4 compares predictions of FOKIT, the three existing sequence-based predictors including C_p (Ma et al. 2007),

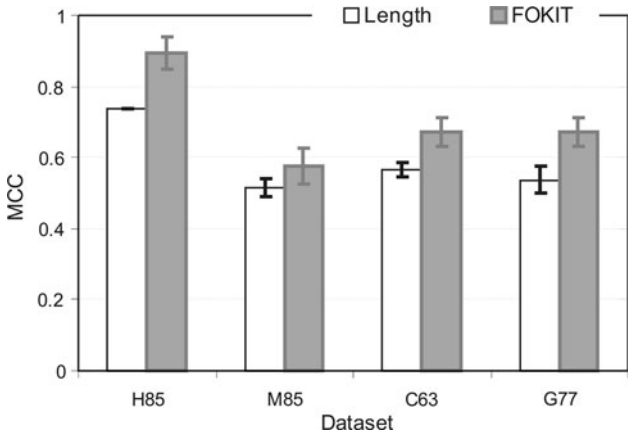

Fig. 2 Comparison of the prediction performance of FOKIT and the predictions based on the chain length for the H85, M85, C63, and G77 datasets. The MCC values (*y* axis) are averaged over 10 random runs of 5CV. Standard deviations are shown using *error bars*

FOLD-RATE Q (Huang and Gromiha 2008) and length-based method (Huang and Cheng 2008), and one structure-based method K-Fold (Capriotti and Casadio 2007) based on jackknife test and 5CV tests on four benchmark datasets. The four competing methods were designed and tested using a single dataset (C_p on the M85, FOLD-RATE Q on the G77, length-based on the H85, and K-Fold on the C63 dataset), while FOKIT is tested on all four datasets. The results show that FOKIT outperforms length-based and K-Fold predictors, i.e., it improves the corresponding MCC by 0.12 and 0.7, respectively. Only the sensitivity of the length-based method is slightly higher than that of FOKIT. Comparison with C_p and FOLD-RATE Q shows that FOKIT provides better sensitivity, while the other two methods give higher specificity. Overall, the MCC/accuracy of these two methods is higher than that of FOKIT. A likely explanation is that C_p and FOLD-RATE Q were designed (parameterized and performed feature selection) and tested using the same M85 and G77 datasets, respectively, while the features used in FOKIT were selected based on the H85 dataset and this method was tested on a different dataset. We compare the C_p method against FOKIT on the H85, C63, and G77 datasets when the latter predictor is also trained on the M85 dataset. The C_p is a linear regression method that takes the chain length and C_{sum} (i.e., the composition of Cys, His, Leu, and Arg) as its inputs, i.e., $C_p = 0.000199 \times \text{Length} + 0.257186 \times C_{sum} - 0.061498$. If $C_p > 0$ then the chain is assumed to be multi-state; otherwise, two-state type is predicted. The results in Table 5 demonstrate that FOKIT outperforms the C_p method on all considered quality indices and test datasets.

Table 5 Comparison of prediction performance between FOKIT and C_p method which were trained on the M85 dataset

Method	Test set	Sensitivity	Specificity	Accuracy	MCC
C_p	H85	0.700	0.840	0.741	0.494
FOKIT		0.817	0.880	0.835	0.651
C_p	C63	0.868	0.800	0.841	0.668
FOKIT		0.895	0.840	0.873	0.735
C_p	G77	0.784	0.731	0.766	0.500
FOKIT		0.843	0.846	0.844	0.668

Table 4 Comparison of the prediction performance between FOKIT and the existing methods based on Jackknife test and fivefold cross validation

Dataset	Method	Jackknife				Fivefold CV (average of 10 runs)			
		Sensitivity	Specificity	Accuracy	MCC	Sensitivity	Specificity	Accuracy	MCC
H85	Length[a]	NA	NA	NA	NA	0.983	0.72	0.906	0.774
	FOKIT	**0.983**	**0.920**	**0.965**	**0.914**	**0.975** ± 0.023	**0.912** ± 0.017	**0.956** ± 0.018	**0.895** ± 0.044
M85	C_p[b]	*0.797*	*0.820*	*0.809*	NA	NA	NA	NA	NA
	FOKIT[c]	**0.837**	**0.738**	**0.788**	**0.579**	**0.846** ± 0.029	**0.724** ± 0.023	**0.786** ± 0.024	**0.576** ± 0.049
		0.836 ± *0.011*	***0.730*** ± *0.022*	***0.783*** ± *0.011*	***0.569*** ± *0.021*				
C63	K-Fold	NA	NA	NA	NA	0.868	0.720	0.810	0.600
	FOKIT	**0.895**	**0.800**	**0.857**	**0.700**	**0.876** ± 0.022	**0.792** ± 0.032	**0.843** ± 0.019	**0.671** ± 0.040
G77	FOLD-RATE Q	0.882	0.923	0.896	0.781	NA	NA	NA	NA
	FOKIT	**0.922**	**0.769**	**0.870**	**0.705**	**0.902** ± 0.016	**0.761** ± 0.030	**0.854** ± 0.017	**0.672** ± 0.039

The CV tests were based on ten runs and the averages and the standard deviations are shown. Results of FOKIT are shown in bold and "NA" denotes results that were not reported by the authors and which could not be duplicated

[a] Huang and Cheng (2008) reported the best result based on 5CV by using the chain length as the only input

[b] Ma et al. (2007) reported the results generated by 1,000 rounds jackknife tests (JKT^v) which are shown in italic. The "accuracy" in Ma et al. (2007) is defined as (Sensitivity + Specificity)/2

[c] Results of JKT^v are shown in italic. We run JKT^v ten times and we report the average, the standard deviations, and the same "accuracy" as in Ma et al. (2007)

Table 6 Predictions FOKIT, C_p, FOLD-RATE Q, and K-Fold on the S17 dataset

PDB id	Source dataset	Chain length	Actual folding type	Predicted folding type			
				FOKIT	C_p	K-Fold	FOLD-RATE Q
1L8WA	M85	348	TS	MS	MS	MS	TS
1B9C	M85	236	MS	MS	MS	MS	TS
1MXI	M85	160	MS	MS	MS	MS	TS
1I1B	M85	153	MS	MS	MS	MS	TS
1JOO	M85	149	MS	TS	MS	MS	TS
1MZK	M85	139	MS	MS	MS	MS	TS
1AZU	M85	128	TS	MS	MS	TS	TS
1ADW	M85	123	MS	TS	TS	MS	TS
1B11	M85	113	MS	MS	MS	TS	TS
1UZC	M85	71	MS	TS	TS	TS	TS
1DTV	M85	67	MS	TS	MS	TS	TS
1L2YA	C63	20	TS	TS	TS	TS	TS
1PGB_	C63	56	TS	TS	TS	TS	TS
1LOP	G77	164	TS	TS	TS	MS	TS
1PIN	G77	153	TS	TS	MS	MS	TS
2HQI	G77	72	TS	TS	TS	TS	TS
1HX5	G77	82	MS	TS	TS	TS	TS

Predictions on the S17 dataset

The S17 dataset was used to validate the quality of predictions for chains that share low, below 25%, identity with the chains used to develop the model, i.e., the H85 dataset. Table 6 compares predictions of FOKIT, C_p, FOLD-RATE Q, and K-Fold. The results of K-Fold and FOLD-RATE Q were derived from the corresponding web servers, the predictions of C_p are computed using the linear model, and FOKIT was trained on the H85 dataset. FOKIT provides better or comparable predictive performances although the other four methods were derived using datasets that overlap with the S17 datasets, i.e., the M85, C63, and G77 datasets that were used to train C_p, K-Fold, and FOLD-RATE Q methods, respectively, share 11, 2, and 4 chains with the S17 dataset. When trained on the H85 dataset, FOKIT obtains sensitivity, specificity, accuracy, and MCC equal 0.714, 0.500, 0.588, and 0.214, respectively. To compare, the MCC values of C_p, FOLD-RATE Q, and K-Fold equal 0.271, 0.169, and 0. The only better result is achieved by C_p, but this is likely due to the substantial overlap between the S17 and M85 datasets (also, Table 5 shows that FOKIT outperforms the C_p method). Table 7 lists the results of FOKIT trained on the M85, C63, and G77 datasets and tested on the S17 dataset to simulate a "fair" comparison, i.e., using the same training and test sets, with the other three methods. The Table shows that the proposed predictor outperforms the three existing methods.

The results suggest that the proposed method provides favorable results when used on novel sequences (i.e., chains that share low, <25%, similarity with the chains used to build the model), which is an important advantage given the limited number of chains with known folding kinetic types.

Table 7 Comparisons of FOKIT with the C_p, K-Fold and FOLD-RATE Q predictors where the methods are trained on the same datasets and tested on the S17 dataset

Method	Training dataset	Sensitivity	Specificity	Accuracy	MCC
FOKIT	M85	0.571	0.800	0.706	0.383
C_p		0.571	0.700	0.647	0.271
FOKIT	C63	0.714	0.700	0.706	0.408
K-Fold		0.571	0.600	0.588	0.169
FOKIT	G77	0.714	0.500	0.588	0.214
FOLD-RATE Q		1.000	0.000	0.412	0.000

Factors governing folding kinetic types

The top-ranked (in the performed feature selection) feature was the chain length, which was previously shown to govern folding types (Huang and Cheng 2008). The effective length (Ivankov and Finkelstein 2004), which was found useful for prediction of folding rates, was not selected. Although these two features are correlated, this suggests that the simple length offers better discriminatory power than the effective length. The other five features include two based on the average RSA values and three AA

Table 8 The mean values of the selected six features and the *P* values that quantify significance of the differences between TS and MS proteins for the H85, M85, C63, and G77 datasets

Dataset	Feature	*P* value	SSD	Mean value	
				TS	MS
H85	Length	9.5E-09	++	79.45 (±20.93)	157.0 (±78.73)
	$AveRSA\_AA_{N}\_SS_{E}$	0.1118	~	0.132 (±0.160)	0.194 (±0.159)
	$AveRSA\_AA_{E}\_SS_{E}$	0.8466	~	0.230 (±0.197)	0.239 (±0.170)
	$AAC_{E}\_SS_{E}$	0.1080	~	0.044 (±0.045)	0.063 (±0.052)
	$AAC_{L}\_Bu_{0.1}$	0.1436	~	0.199 (±0.135)	0.234 (±0.069)
	$AAC_{W}\_Ex_{0.2}$	0.0307	++	0.001 (±0.004)	0.005 (±0.009)
M85	Length	5.4E-05	++	85.95 (±48.44)	138.88 (±63.16)
	$AveRSA\_AA_{N}\_SS_{E}$	0.1535	~	0.122 (±0.139)	0.170 (±0.165)
	$AveRSA\_AA_{E}\_SS_{E}$	0.8377	~	0.239 (±0.191)	0.230 (±0.178)
	$AAC_{E}\_SS_{E}$	0.3490	~	0.042 (±0.038)	0.051 (±0.046)
	$AAC_{L}\_Bu_{0.1}$	0.0030	++	0.174 (±0.131)	0.238 (±0.088)
	$AAC_{W}\_Ex_{0.2}$	0.0016	++	0.000 (±0.001)	0.004 (±0.008)
C63	Length	0.0003	++	79.76 (±44.91)	141.8 (±67.65)
	$AveRSA\_AA_{N}\_SS_{E}$	0.1706	~	0.131 (±0.141)	0.189 (±0.168)
	$AveRSA\_AA_{E}\_SS_{E}$	0.4986	~	0.231 (±0.197)	0.264 (±0.172)
	$AAC_{E}\_SS_{E}$	0.0637	+	0.040 (±0.039)	0.065 (±0.051)
	$AAC_{L}\_Bu_{0.1}$	0.0022	++	0.168 (±0.145)	0.248 (±0.075)
	$AAC_{W}\_Ex_{0.2}$	0.0226	++	0.000 (±0.002)	0.004 (±0.008)
G77	Length	0.0003	++	83.65 (±35.76)	140.7 (±66.43)
	$AveRSA\_AA_{N}\_SS_{E}$	0.2123	~	0.128 (±0.145)	0.178 (±0.166)
	$AveRSA\_AA_{E}\_SS_{E}$	0.9959	~	0.258 (±0.195)	0.258 (±0.175)
	$AAC_{E}\_SS_{E}$	0.1419	~	0.045 (±0.039)	0.063 (±0.051)
	$AAC_{L}\_Bu_{0.1}$	0.0543	+	0.189 (±0.122)	0.229 (±0.082)
	$AAC_{W}\_Ex_{0.2}$	0.0715	+	0.001 (±0.005)	0.005 (±0.008)

The SSD column indicates features that are characterized by statistically significant differences at 0.05 level (++), at 0.1 level (+) and no significant differences (~)

composition-based features. We investigate statistical significance of the differences of the values of these features between the TS and the MS folders on the four datasets. Table 8 gives the *P* values of two-sided *t* tests (when both distributions are normal) or two-sided Wilcoxon test (when at least one of the distributions is not normal). The normality was tested using Shapiro–Wilk test at the 0.05 significance.

Although as expected the chain length is significantly different between the two protein sets on all four datasets, the $AAC_{W}\_Ex_{0.2}$ feature is also shown to provide statistically significant discrimination. This feature quantifies the content of solvent exposed (at the 0.2 cutoff) Trp in the protein chain. The mean values of the $AAC_{W}\_Ex_{0.2}$ feature for the TS proteins are significantly smaller than the values for the MS proteins. The proteins that contain no or a very few exposed Trp residues, which are hydrophobic, are more likely to utilize TS folding process, while increased content of exposed Trp correlates with the MS folders. Figure 3 shows that the difference disappears if we consider Trp residues irrespective of their solvent exposure. The mean values of AAC_{W} for the TS and the MS proteins in the H85 dataset equal 0.016 and 0.017, respectively. The Trp residues have been implicated in long-range interaction that stabilize the protein structure (Zhou et al. 2007; Klein-Seetharaman et al. 2002), and their mutation was shown to increase refolding rate and decrease stability of the native fold (Linke et al. 2004). Our finding is also supported by several experimental studies. Zhou et al. (2007) demonstrated that a single mutation (W62G) of the surface Trp in hen egg-white lysozyme resulted in a less stable structure than that of the wild type. Linke et al. (2004) showed that the mutation of a solvent exposed Trp to Phe in OEP16 protein resulted in speeding up the folding process. We use the work of Calloni et al. (2003) who suggested that the average hydrophobicity of a protein sequence is an important determinant of its folding rate by comparing the folding mechanics of the N-terminal domain of HypF from *Escherichia coli* (HypF-N) and two extensively studied human proteins, muscle and common-type

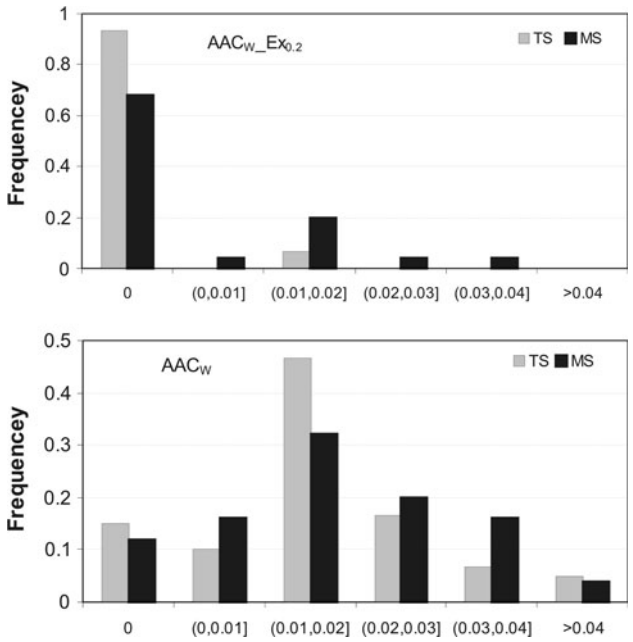

Fig. 3 Distributions of values of features $AAC_W\_Ex_{0.2}$ (**a**) and AAC_W (**b**) among the TS and MS folders in the H85 dataset

acylphosphatases (mAcP and ctAcP), as our case studies. These proteins have the same topology ($\beta\alpha\beta\beta\alpha\beta$ motif), but HypF-N shares only 22 and 26% sequence identity with mAcP and ctAcP (Calloni et al. 2003), respectively. The mAcP and ctAcP proteins have been shown to fold with the two-state kinetic pathway, while HypF-N utilizes the three-state folding, i.e., it collapses into a partially folded intermediate before reaching the fully folded conformation. Furthermore, HypF-N was found to fold rapidly with a rate constant that is approximately two and three orders of magnitude faster than ctAcP and mAcP, respectively. This suggests that the structural similarity may not imply the similarity in folding kinetics or rate. These proteins include two evolutionary conserved Trp residues. The RSA values of these Trp residues computed with DSSP (Kabsch and Sander 1983) are 38.3% for Trp27 and 5% for Trp81 in HypF-N (PDB ID: 1GXT), 11.2% for Trp38 and 4.6% for Trp64 in mAcP (PDB ID: 1APS), and 19.5% for Trp38, and 7.5% for Trp64 in ctAcP (PDB ID: 2ACY), respectively; see Fig. 4. Only the Trp27 in HypF-N is exposed with respect to the RSA cutoff of 20%. This means that the value of the $AAC_W\_Ex_{0.2}$ feature is greater than zero for HypF-N and equal zero for both mAcP and ctAcP, which supports our observation that exposure of Trp could be an important marker for MS kinetic type.

Table 8 also shows that $AAC_L\_Bu_{0.1}$, which quantifies content of buried Leu (using cutoff of 0.1), also provides significant discrimination for three out of the four datasets. Dong et al. (2008) found that buried hydrophobic residues, including Leu, contribute to the stabilization of the structure. Their experiments reveal that mutants of Ribonuclease HII from hyperthermophile *Thermococcus kodakaraensis* (Tk-RNase HII), which substitute large buried hydrophobic residues by smaller residues (Leu/Ile to Ala), unfold faster than the wild-type protein. Although it is not a direct evidence for differences in folding kinetic type, we hypothesize that the increase of the content of the buried Leu residues is a marker of the MS proteins. A number of small single-domain proteins have been shown to form intermediates during folding (Park et al. 1999; Ferguson et al. 1999; Laurents et al. 2000). The E colicin binding immunity proteins (Im7 and Im9) are appropriate for investigating the kinetic mechanics of single-domain proteins as we used them as our case studies. They have 87 and 86 residues, respectively, share 60% sequence identity, and fold to the same native structure with four helices but by different kinetic types (Ferguson et al. 1999; Friel et al. 2004). At the same conditions with 7.0 pH and 10°C, Im9 folds with TS process, while the less stable homologue, Im7, folds with three-state pathway. Given that V19, V37, and V71 residues in Im9* (histidine-tagged Im9) are replaced with the

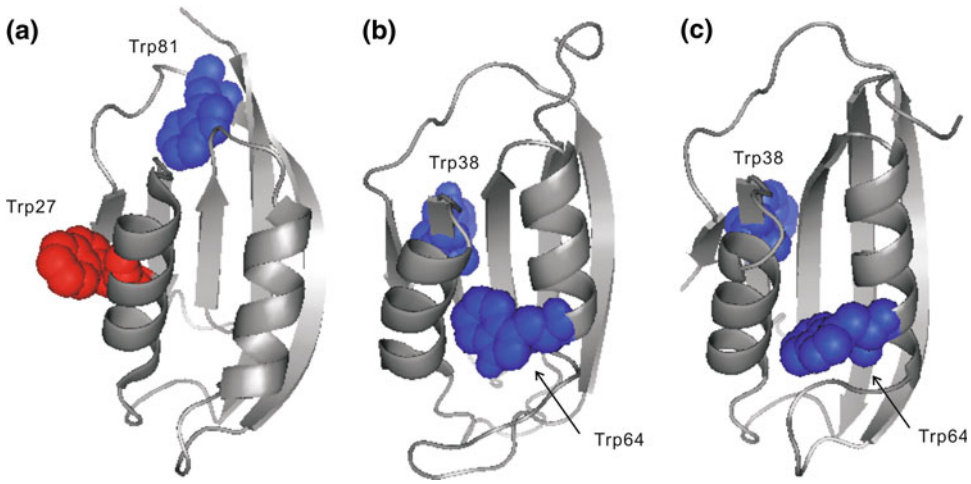

Fig. 4 Three-dimensional cartoon structures of **a** HypF-N (PDB ID: 1GXT), **b** mAcP (PDB ID: 1APS), and **c** ctAcP (PDB ID: 2ACY). The Trp residues are shown using filled spheres where exposed Trp residues (using RSA cutoff of 20%) are in *red* and buried Trp residues are in *blue*

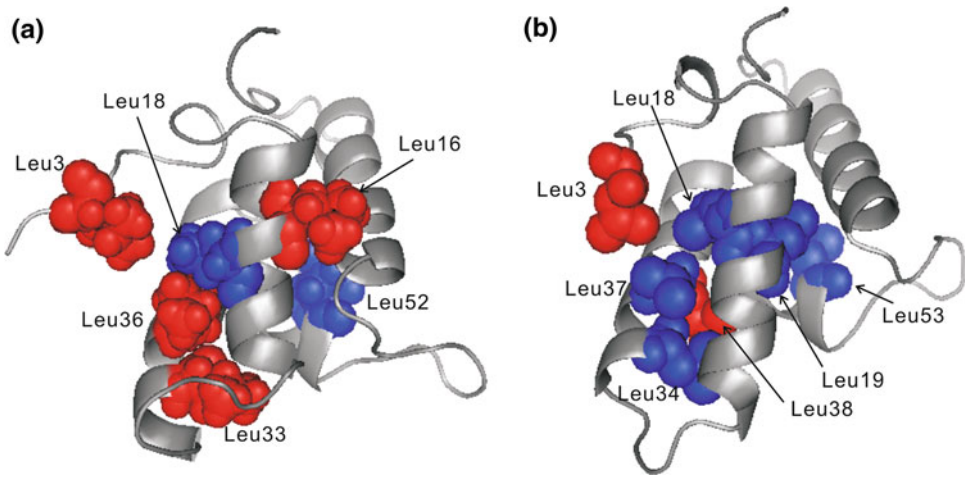

Fig. 5 Three-dimensional cartoon structures of **a** Im9 (PDB ID: 1IMQ) and **b** Im7 (PDB ID: 1CEI). The Leu residues are shown using *filled spheres* where exposed Leu residues (using RSA cutoff of 10%) are in *red* and buried Leu residues are in *blue*

equivalent residues in Im7* (histidine-tagged Im7), the mutated chains (V19L, V37L and V71I) of Im9* were found by Friel et al. to switch from TS to three-state folding process (Friel et al. 2004). In each mutation, the Val residue in Im9 is less hydrophobic than the equivalent residues (Leu/Ile) in Im7 and this increase in the hydrophobicity may explain the switching. We examined the solvent exposure the Leu residues in both Im9 and Im7, see Fig. 5. There are six Leu residues in Im9 at positions 3, 16, 18, 33, 36, and 52 and their RSA values equal 63.9, 32.2, 9.8, 11.5, 19.7, and 9.8%, respectively. The RSA values of Leu residues at the 3, 18, 19, 34, 37, 38, and 53 positions in Im7 are 71.5, 8.7, 2.2, 6.0, 6.6, 12.0, and 1.1%, respectively. Using the 10% cutoff, two out of six Leu residues are buried in Im9 and five out of seven are buried in Im7. This supports our observation that the larger content of buried Leu residues may be associated with the MS folding.

Table 8 also reveals that the AveRSA_AA$_N$_SS$_E$, AveRSA_AA$_E$_SS$_E$, and AAC$_E$_SS$_E$ features show no statistically significant differences between the TS and the MS proteins. We note that AveRSA_AA$_N$_SS$_E$ and AAC$_E$_SS$_E$ have higher average values for the MS proteins in all four datasets. The lack of significance implies that these are not strong markers when used individually, but they are shown to improve the predictions when used in combination with the other features (see Fig. 1).

Conclusions and discussion

The chain length and the topology of the native fold have been recently shown as important factors determining the folding kinetic types (Capriotti and Casadio 2007; Huang and Cheng 2008). Huang and Cheng (2008) used the length cutoff of 112 to discriminate between the TS and MS proteins. However, a number of small, single-domain proteins have been shown to fold utilizing the MS process. In addition, the structural similarity does not imply the similarity in folding kinetic types as shown in the case studies concerning HypF-N, mAcP and ctAcP, and for Im9 and Im7. We utilized a wrapper-based feature selection to find a small set of complementary features which hybridize information concerning AA composition and predicted secondary structure and solvent accessibility. We demonstrated that inclusion of solvent exposure helps in discrimination of the folding kinetic types. Some of the considered features are shown to be strong markers for the folding kinetic types, i.e., they provide statistically significant differences between the TS and MS folders. They are sensitive to presence of individual exposed/buries residues and thus they could quantify the effect of certain mutations that can switch the folding type, as shown using the case studies. More specifically, we found that the increased content of exposed Trp and buried Leu are indicative of the MS folding process, which implies that the exposure/burial of certain hydrophobic residues may play important role in the formation of folding intermediates.

Acknowledgments The authors would like to thank Emidio Capriotti, Michael Gromiha, Ji-Tao Huang, and Bin-Guang Ma for providing their datasets. This work was supported in part by the National Natural Science Foundation of China (Grant No. 61003187), Zhejiang Provincial Natural Science Foundation of China (Grant No. Y1090688), and NSERC Canada.

References

Ahmad S, Gromiha MM, Sarai A (2003) Real value prediction of solvent accessibility from amino acid sequence. Proteins 50:629–635

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Anfinsen CB (1973) Principles that govern the folding of protein chains. Science 181:223–230

Bogatyreva NS, Osypov AA, Ivankov DN (2009) KineticDB: a database of protein folding kinetics. Nucleic Acids Res 37:D342–D346

Borgia A, Bonivento D, Travaglini-Allocatelli C, Di Matteo A, Brunori M (2006) Unveiling a hidden folding intermediate in *c*-type cytochromes by protein engineering. J Biol Chem 281:9331–9336

Bryson K, McGuffin LJ, Marsden RL, Ward JJ, Sodhi JS, Jones DT (2005) Protein structure prediction servers at University College London. Nucleic Acids Res 33:W36–W38

Callender RH, Dyer RB, Gilmanshin R, Woodruff WH (1998) Fast events in protein folding: the time evolution of primary processes. Annu Rev Phys Chem 49:173–202

Calloni G, Taddei N, Plaxco KW, Ramponi G, Stefani M, Chiti F (2003) Comparison of the folding processes of distantly related proteins. Importance of hydrophobic content in folding. J Mol Biol 330:577–591

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

Chen K, Kurgan LA (2007) PFRES: protein fold classification by using evolutionary information and predicted secondary structure. Bioinformatics 23(21):2843–2850

Chen K, Kurgan LA, Ruan J (2007) Prediction of flexible/rigid regions in proteins from sequences using collocated amino acid pairs. BMC Struct Biol 7:25

Cranz-Mileva S, Friel CT, Radford SE (2005) Helix stability and hydrophobicity in the folding mechanism of the bacterial immunity protein Im9. Protein Eng Des Sel 18:41–50

Cronin MT, Aptula AO, Dearden JC, Duffy JC, Netzeva TI, Patel H, Rowe PH, Schultz TW, Worth AP, Voutzoulidis K, Schüürmann G (2002) Structure-based classification of antibacterial activity. J Chem Inf Comput Sci 42:869–878

Dill KA, Ozkan SB, Shell MS, Weikl TR (2008) The protein folding problem. Annu Rev Biophys 37:289–316

Dong H, Mukaiyama A, Tadokoro T, Koga Y, Takano K, Kanaya S (2008) Hydrophobic effect on the stability and folding of a hyperthermophilic protein. J Mol Biol 378:264–272

Dor O, Zhou Y (2007) Real-SPINE: an integrated system of neural networks for real-value prediction of protein structural properties. Proteins 68:76–81

Esposito G, Ricagno S, Corazza A, Rennella E, Gümral D, Mimmi MC, Betto E, Pucillo CE, Fogolari F, Viglino P, Raimondi S, Giorgetti S, Bolognesi B, Merlini G, Stoppini M, Bolognesi M, Bellotti V (2008) The controlling roles of Trp60 and Trp95 in beta2-microglobulin function, folding and amyloid aggregation properties. J Mol Biol 378:887–897

Feng H, Zhou Z, Bai Y (2005) A protein folding pathway with multiple folding intermediates at atomic resolution. Proc Natl Acad Sci USA 102:5026–5031

Ferguson N, Capaldi AP, James R, Kleanthous C, Radford SE (1999) Rapid folding with and without populated intermediates in the homologous four-helix proteins Im7 and Im9. J Mol Biol 286:1597–1608

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

Finkelshtein AV, Galzitskaya OV (2004) Physics of protein folding. Phys Life Rev 1:23–56

Friel CT, Beddard GS, Radford SE (2004) Switching two-state to three-state kinetics in the helical protein Im9 via the optimization of stabilizing non-native interactions by design. J Mol Biol 342:261–273

Fulton KF, Bate MA, Faux NG, Mahmood K, Betts C, Buckle AM (2007) Protein Folding Database (PFD 2.0): an online environment for the International Foldeomics Consortium. Nucleic Acids Res 35:D304–D307

Galzitskaya OV, Garbuzynskiy SO, Ivankov DN, Finkelstein AV (2003) Chain length is the main determinant of the folding rate for proteins with three-state folding kinetics. Proteins 51:162–166

Galzitskaya OV, Bogatyreva NS, Ivankov DN (2008a) Compactness determines protein folding type. J Bioinform Comput Biol 6:667–680

Galzitskaya OV, Danielle C, Reifsnyder DC, Bogatyreva NS, Ivankov DN, Garbuzynskiy SO (2008b) More compact protein globules exhibit slower folding rates. Proteins 70:329–332

Gong H, Isom DG, Srinivasan R, Rose GD (2003) Local secondary structure content predicts folding rates for simple, two-state folding proteins. J Mol Biol 327:1149–1154

Gromiha MM (2009) Multiple contact network is a key determinant to protein folding rates. J Chem Inf Model 49:1130–1135

Gromiha MM, Selvaraj S (2001) Comparison between long-range interactions and contact order in determining the folding rate of two-state proteins: application of long-range order to folding rate prediction. J Mol Biol 310:27–32

Hsu CW, Lin CJ (2002) A comparison on methods for multi-class support vector machines. IEEE Trans Neural Netw 13:415–425

Huang JT, Cheng JP (2008) Differentiation between two-state and multi-state folding proteins based on sequence. Proteins 72:44–49

Huang LT, Gromiha MM (2008) Analysis and prediction of protein folding rates using quadratic response surface models. J Comp Chem 29:1675–1683

Huang JT, Cheng JP, Chen H (2007) Secondary structure length as a determinant of folding rate of proteins with two- and three-state kinetics. Proteins 67:12–17

Inaba K, Kobayashi N, Fersht AR (2000) Conversion of two-state to multi-state folding kinetics on fusion of two protein foldons. J Mol Biol 302:219–233

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

Ivankov DN, Garbuzynskiy SO, Alm E, Plaxco K, Baker D, Finkelstein AV (2003) Contact order revisited: influence of protein size on the folding rate. Protein Sci 12:2057–2062

Ivankov DN, Bogatyreva NS, Lobanov MY, Galzitskaya OV (2009) Coupling between properties of the protein shape and the rate of protein folding. PLoS One 4:e6476

Jackson SE (1998) How do small single-domain proteins fold? Fold Des 3:R81–R91

Jiang Y, Iglinski P, Kurgan L (2009) Prediction of protein folding rates from primary sequences using hybrid sequence representation. J Comput Chem 30:772–783

Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637

Kamagata K, Arai M, Kuwajima K (2004) Unification of the folding mechanisms of non-two-state and two-state proteins. J Mol Biol 339:951–965

Klein-Seetharaman J, Oikawa M, Grimshaw SB, Wirmer J, Duchardt E, Ueda T, Imoto T, Smith LJ, Dobson CM, Schwalbe H (2002) Long-range interactions within a nonnative protein. Science 295:1719–1722

Kohavi R, John GH (1997) Wrappers for feature subset selection. Artif Intell 97:273–324

Krishna MM, Hoang L, Lin Y, Englander SW (2004) Hydrogen exchange methods to study protein folding. Methods 34:51–64

Laurents DV, Corrales S, Elias-Arnanz M, Sevilla P, Rico M, Padmanabhan S (2000) Folding kinetics of Phage 434 Cro proteins. Biochemistry 39:13963–13973

Lindorff-Larsen K, Best RB, Depristo MA, Dobson CM, Vendruscolo M (2005) Simultaneous determination of protein structure and dynamics. Nature 433:128–132

Linke D, Frank J, Pope MS, Soll J, Ilkavets I, Fromme P, Burstein EA, Reshetnyak YK, Emelyanenko VI (2004) Folding kinetics and structure of OEP16. Biophys J 86:1479–1487

Ma BG, Guo JX, Zhang HY (2006) Direct correlation between proteins' folding rates and their amino acid compositions: an ab initio folding rate prediction. Proteins 65:362–372

Ma BG, Chen LL, Zhang HY (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

Maity H, Maity M, Krishna MM, Mayne L, Englander SW (2005) Protein folding: the stepwise assembly of foldon units. Proc Natl Acad Sci USA 102:4741–4746

Matthews BW (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 405:442–451

Ouyang Z, Liang J (2008) Predicting protein folding rates from geometric contact and amino acid sequence. Protein Sci 17:1256–1263

Park SH, Shastry MC, Roder H (1999) Folding dynamics of the B1 domain of protein G explored by ultrarapid mixing. Nature Struct Biol 6:943–947

Plaxco KW, Simons KT, Baker D (1998) Contact order, transition state placement and the refolding rates of single domain proteins. J Mol Biol 277:985–994

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

Ricagno S, Raimondi S, Giorgetti S, Bellotti V, Bolognesi M (2009) Human beta-2 microglobulin W60 V mutant structure: Implications for stability and amyloid aggregation. Biochem Biophys Res Commun 380:543–547

Scheraga HA, Khalili M, Liwo A (2007) Protein-folding dynamics: overview of molecular simulation techniques. Annu Rev Phys Chem 58:57–83

Schuler B, Lipman EA, Eaton WA (2002) Probing the free-energy surface for protein folding with single-molecule fluorescence spectroscopy. Nature 419:743–747

Shen HB, Song JN, Chou KC (2009) Prediction of protein folding rates from primary sequence by fusing multiple sequential features. J Biomed Sci Eng 2:136–143

Song J, Burrage K (2006) Predicting residue-wise contact orders in proteins by support vector regression. BMC Bioinform 7:425

Sosnick TR, Dothager RS, Krantz BA (2004) Differences in the folding transition state of ubiquitin indicated by φ and ψ analyses. Proc Natl Acad Sci USA 101:17377–17382

Udgaonkar JB (2008) Multiple routes and structural heterogeneity in protein folding. Annu Rev Biophys 37:489–510

Vapnik V (1998) Statistical learning theory. Wiley, New York

Viguera AR, Serrano L (2003) Hydrogenexchange stability analysis of Bergerac-Src homology 3 variants allows the characterization of a folding intermediate in equilibrium. Proc Natl Acad Sci USA 100:5730–5735

Witten IH, Frank E (2005) Data mining: practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann, San Francisco

Yu L, Liu H (2003) Feature selection for high-dimensional data: a fast correlation-based filter solution. In: Proceedings of the 10th international conference on machine learning, pp 856–863

Zhang H, Zhang T, Chen K, Shen S, Ruan J, Kurgan LA (2008) Sequence based residue depth prediction using evolutionary information and predicted secondary structure. BMC Bioinform 9:388

Zhang H, Zhang T, Chen K, Shen S, Ruan J, Kurgan L (2009) On the relation between residue flexibility and local solvent accessibility in proteins. Proteins 76:617–636

Zhou R, Eleftheriou M, Royyuru AK, Berne BJ (2007) Destruction of long-range interactions by a single mutation in lysozyme. Proc Natl Acad Sci USA 104:5824–5829