BRIEFINGS IN BIOINFORMATICS. VOL 16. NO 2. 314–324 doi:10.1093/bib/bbu007
Advance Access published on 11 March 2014

Towards more accurate prediction of protein folding rates: a review of the existing web-based bioinformatics approaches

Catherine Ching Han Chang, Beng Ti Tey, Jiangning Song and Ramakrishnan Nagasundara Ramanan

Submitted: 25th September 2013; Received (in revised form): 9th February 2014

Abstract

The understanding of protein-folding mechanisms is often considered to be an important goal that will enable structural biologists to discover the mysterious relationship between the sequence, structure and function of proteins. The ability to predict protein-folding rates without the need for actual experimental work will assist the research work of structural biologists in many ways. Many bioinformatics tools have emerged in the past decade, and each has showcased different features. In this article, we review and compare eight web-based prediction tools that are currently available and that predominantly predict the protein-folding rate. The prediction performance, usability and utility, together with the prediction tool development and validation methodologies for these tools, are critically reviewed. This article is presented in a comprehensible manner to assist readers in the process of selecting the most appropriate bioinformatics tools to meet their needs.

Keywords: *prediction tool; in silico prediction; machine learning algorithm; prediction model; statistical analysis; molecular biology*

INTRODUCTION

The fate of a protein to be either functional or inactive depends upon the folding mechanism. The failure of a protein to fold into the intended three-dimensional (3D) structure will result in a misfolded protein. The misfolded protein, which generally exists in the form of an inclusion body, is typically insoluble and biologically inactive [1]. Researchers often concentrate on the determination of protein-folding kinetics and rate constants because these are important factors that can contribute to our understanding of the protein-folding mechanism [2]. Experimental determination of folding kinetics and rate constants is generally time-consuming and

Corresponding authors. Ramakrishnan Nagasundara Ramanan, Chemical Engineering Discipline, School of Engineering, Monash University, Jalan Lagoon Selatan, 46150 Bandar Sunway, Selangor, Malaysia. Tel.: +603 5514 6235; Fax: +603 5514 6207. E-mail: ramanan@monash.edu; Jiangning Song, Department of Biochemistry and Molecular Biology, Faculty of Medicine, Monash University, Melbourne, Victoria 3800, Australia. Tel.: +613 9902 9304; Fax: +613 9902 9500. E-mail: Jiangning.Song@monash.edu

Catherine Ching Han Chang received her degree in Chemical Engineering from the Monash University Malaysia. She is currently pursuing her PhD in Chemical Engineering at the Monash University Malaysia. Her research interests include soluble protein expression, mathematical modeling and bioinformatics.

Beng Ti Tey is a Professor of Chemical Engineering Discipline in the School of Engineering at the Monash University Malaysia. His area of expertise includes recombinant protein production and purification.

Jiangning Song is a Senior Research Fellow at the Department of Biochemistry and Molecular Biology, Faculty of Medicine, Monash University Australia. He is also a Principal Investigator at the Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences (CAS). His research interests are bioinformatics, systems biology, machine learning, systems pharmacology and enzyme engineering.

Ramakrishnan Nagasundara Ramanan is a Senior Lecturer of Chemical Engineering Discipline in the School of Engineering at the Monash University Malaysia. His research interests include biomolecular engineering, bioprocess modeling and interaction of biomolecules.

labor-intensive, and accordingly, bioinformatics tools have been developed as alternatives. Interestingly, both protein solubility and folding kinetics are closely linked to one another, because the folding kinetics of proteins determine the tendency of a protein to fold into its soluble native state or to misfold forming an insoluble inclusion body [3, 4]. In this regard, a variety of bioinformatics tools have been developed in the past decade that focus on the prediction of either protein-folding kinetics [5–12] or protein solubility [4, 13–22]. These tools use information regarding the 3D structure of proteins, and their predictions have been shown to correlate closely with experimentally determined folding rates. The features extracted from the 3D structural information include contact order [23–25], long-range order [24, 25], total contact distance [25], *n*-order contact distance [26] and geometric contact number [12]. Although the predictions of protein-folding rates using these aforementioned methods have yielded reasonably good performance in terms of correlation coefficients, prediction based merely on the amino acid sequence is much preferred in practice because the amino acid sequence is the most readily available information. Apart from the simplicity of using the amino acid sequence as the only basis for prediction, the tremendous growth in genomic studies has caused the rate of new protein sequence discovery to outpace protein structural determination [27]. This leaves a large gap between the number of protein sequence entries and the number of proteins for which the structure is known. As a result, this expanding sequence-structure gap makes sequence-based bioinformatics tools increasingly important for researchers. Moreover, sequence-based tools present attractive advantages in handling high-throughput data generated in the fast-growing proteomics field.

Because the code for protein-folding kinetics has been suggested to lie in the amino acid sequence [28, 29], a variety of bioinformatics tools that use the amino acid sequence have been developed to predict protein-folding kinetics. Although most of these tools apply similar algorithms in the development procedure, they differ greatly in terms of performance, usability and utility, which subsequently affect the analysis outcome. Consequently, non-bioinformaticians are often unable to fully explore these bioinformatics tools owing to a lack of familiarity with and fundamental knowledge of bioinformatics tools. Gromiha and Huang in their work have previously reviewed the existing machine learning algorithms that predict both the protein-folding rate and stability of mutant proteins [8]. These authors have thoroughly reviewed the folding parameters that will possibly affect the protein-folding process as well as the development progress of structure-based parameters. In addition, they included a short introduction discussing two databases and five web-based prediction tools that can predict protein-folding rates. Distinctively, the present work emphasizes the usability, utility and performance of the available prediction tools as well as the algorithms used. This comparative review is intended to enable readers to compare the available web-based tools and to assist in decisionmaking.

This article presents a comprehensive comparison of eight web-based tools for protein-folding rate predictions. The performance, usability and utility of these tools, as well as the algorithm adopted in the development stage, are critically reviewed to serve as a gateway to assist researchers, especially non-bioinformaticians, in choosing the most suitable prediction tool to meet different requirements. In addition, a brief description of the alternative tools for protein-folding rate prediction is provided. As the number of proteins discovered each year continues to grow, protein-folding rate prediction tools appear to have a promising future because the information acquired is valuable for assisting researchers in studying protein-folding mechanisms. The ability to comprehend and reveal the relationship between protein sequence and the corresponding-folding rate will benefit the field of pathology because some misfolded proteins have been generally considered to be responsible for neurodegenerative diseases, such as Alzheimer's disease [1, 30].

EXISTING PROTEIN-FOLDING RATE PREDICTION TOOLS

The comparison between the web-based tools currently available for protein-folding rate prediction is summarized in Table 1. Tools assessed in this study include SFoldRate [12], FOLD-RATE [31], Pred-PFR [10], FoldRate [6], K-Fold [7], PRORATE [9], SWFoldRate [5] and SeqRate [11]. These are the only web-based protein-folding rate prediction tools that are freely accessible to researchers at academic institutions. The criteria used in the comparison will be discussed in depth in the following sections.

Table 1: A comparison of the protein-folding rate prediction tools reviewed[a]

Tool[b]	SFoldRate	FOLD-RATE	Pred-PFR	FoldRate	K-Fold	PRORATE	SWFoldRate	SeqRate
Correlation coefficient	0.82	0.87[c]	0.88	0.88	0.74	0.85[d]; 0.88[e]	0.93	0.81[d];0.80[e]
Statistical deviation[f]	n/a	$x \pm 1.19$[g] (MAD); $x \pm 0.23$[h] (MAD)	$x \pm 2.03$	$x \pm 2.03$	$x \pm 1.2$ (SE); $x \pm 0.75$[i] (MAD)	$x \pm 1.95$[d,g]; $x \pm 2.12$[e,g]; $x \pm 1.34$[d,h]; $x \pm 1.77$[e,h]	$x \pm 2.27$ (SE)	$x \pm 0.79$[d] (MAD); $x \pm 0.68$[e] (MAD)
Application type	WB	WB	WB	WB	WB	WB	WB	WB, SA
Development method	Statistical method with multiple linear regressions	Statistical method with multiple linear regressions	Ensemble predictor with multiple statistical models	Ensemble predictor with multiple statistical models	SVM classifier with linear kernel	SVR with polynomial kernel	Non-linear SVM regression model with sliding window	Non-linear SVM classifier with radial basis Gaussian kernel and regression model
Performance Evaluation strategy	Leave-one-out cross-validation	Leave-one-out cross-validation and back-check prediction	Leave-one-out cross-validation	Leave-one-out cross-validation	Cross-validation	Leave-one-out cross-validation and back-check prediction	Leave-one-out cross-validation	Independent test
Experimental verification	No	No	No	No	No	No	No	Yes
Size of training data set (number of proteins)	80	77	80	80	63	80	79	54
Size of test data set (number of proteins)	n/a	n/a	n/a	n/a	n/a	n/a	n/a	7
Input format	FASTA	Plain sequence	Plain sequence	Plain sequence	PDB code/file	PDB file	FASTA	Plain sequence
Additional input	No	Structural class of protein	No	No	No	Protein-folding kinetic state	No	Protein-folding kinetic state
Outputs	Folding rate in natural logarithm	Residues composition, protein structural class and folding rate in natural logarithm	Folding rate in natural logarithm	Folding rate in natural logarithm, half-folding time	Contact order, reliability index, protein-folding kinetic state and folding rate in logarithm based 10	Topology parameters, network parameters, protein-folding kinetic state and folding rate in natural logarithm	Folding rate in natural logarithm	Protein-folding kinetic state, contact number, contact order and folding rate in logarithm based 10 and natural logarithm
References	[12]	[31]	[10]	[6]	[7]	[9]	[5]	[11]

[a]n/a = not applicable; WB = web-based; SA = stand-alone; and SVR = support vector regression. [b]All prediction tools provide open access and the URL addresses to access the different prediction tools are as follows: SFoldRate - http://gila.bioengr.uic.edu/lab/tools/foldingrate/fr0.html; FOLD-RATE - http://psfs.cbrc.jp/fold-rate/; Pred-PFR - http://www.csbio.sjtu.edu.cn/bioinf/FoldingRate/#; FoldRate - http://www.csbio.sjtu.edu.cn/bioinf/FoldRate/; K-Fold - http://gpcr.biocomp.unibo.it/cgi/predictors/K-Fold/K-Fold.cgi; PRORATE - http://sunflower.kuicr.kyoto-u.ac.jp/~sjn/folding/webserver.html; SWFoldRate - http://www.jci-bioinfo.cn/swfrate/input.jsp; SeqRate - http://casp.rnet.missouri.edu/fold_rate/index.html. [c]Outcome with respect to proteins of unknown structural class. [d]Outcome with respect to two-state proteins. [e]Outcome with respect to multistate proteins. [f]x – notation for predicted folding rate from respective prediction tool; statistical deviation reported as root mean square error unless specified in parentheses. [g]Outcome obtained using LOOCV. [h]Outcome obtained using back-check prediction. [i]Outcome reported in other literature [11].

Prediction tool development method

With respect to the aforementioned prediction tools, both statistical and machine learning algorithms have been used as the basal algorithms for developing the prediction tools. In general, statistical algorithms include linear regression and logistic regression, whereas machine learning algorithms include decision trees and neural nets [32]. Statistical algorithms such as linear regression often produce a straightforward and comprehensible representation to relate the input variables to the corresponding output [32]. Multiple linear regression can be used to take into account the effect of different input variables. In this regard, SFoldRate and FOLD-RATE are built based on multiple linear regression. Different weights are assigned to indicate the significance of respective features, which have been previously found to be highly correlated with protein-folding rate.

Seven individual predictors are integrated into one ultimate predictor, namely, Pred-PFR, to conduct the protein-folding rate prediction. The ensemble of multiple individual predictors, with each functioning based on its own special features, has been proven to perform more effectively when applied to sophisticated biological systems [33, 34]. Similarly, FoldRate adopts an ensemble predictor that fuses three individual predictors based on different statistical models. K-Fold and PRORATE are developed using machine learning algorithms, particularly support vector machine (SVM). PRORATE is developed using support vector regression, which is considered to be one of the SVM categories. Linear kernel functions are incorporated into the SVMs of both K-Fold and PRORATE. The incorporation of kernel functions enables the application of SVM to solve biological modeling problems, which often involve the processing of non-vector data, such as nucleotide and protein sequences [35, 36].

Machine learning algorithms are capable of enhancing the performance of statistical methods by introducing automated information discovery and processing techniques to the existing statistical methods. Both statistical and machine learning algorithms are complementary to each other, rather than incompatible. Statistical techniques provide estimates on the probability of the possible outcome, while machine learning algorithms often specify a deterministic result for a classification task [37]. Moreover, statistical models are exceptionally suitable for processing continuous attributes, which subsequently provide interpolative or extrapolative approximations [38]. Conversely, machine learning algorithms are frequently used to analyze discrete attributes and give only predictive ranges [38, 39]. In view of this, SWFoldRate and SeqRate represent prediction tools that incorporate both statistical and machine learning techniques. This is for the purpose of producing a powerful prediction tool that makes use of the advantages of both techniques while offsetting their respective shortcomings. Therefore, both SWFoldRate and SeqRate can be seen as tools that have been developed using the most robust methodology out of the eight tools reviewed in this article.

Feature selection

For any prediction tool, it is essential to perform systematic feature selection procedures. This is aimed at reducing the risk of overlooking certain features that perchance are equally influential in determining folding rates. Inclusion of too many features may give rise to an over-fitting issue, while inclusion of insufficient and redundant features may not result in the best-performing prediction tool [5, 35]. PRORATE adopts a recursive elimination strategy during feature selection, aiming to improve the prediction performance by removing features with insignificant influence on the prediction accuracy [9]. The optimal features in SWFoldRate are selected using combined forward feature selection and sequential backward selection methods. Sequential forward and backward selection [40] is one of the classic deterministic heuristic feature subset selection algorithms [41]. Heuristic search outperforms an exhaustive search owing to reduced computational effort, making it more practical, especially when handling a large pool of features [41]. In contrast, the knowledge-based approach has been applied in cases where a function needs to be built to map the input to protein folding rates [11]. The features that yield the best correlation scores will be adopted in the training process of the respective prediction tools. This method is used in SeqRate along with the leave-one-out cross-validation (LOOCV) procedure. Similarly, FOLD-RATE has adopted a similar strategy to select the optimal combination of features among the 49 features extracted at the initial stage. Prediction tools that failed to undergo a thorough and systematic feature selection procedure would generally result in reduced performance [41]. Accordingly, this places FoldRate, Pred-PFR and SFoldRate in an unfavorable position, as their

feature selection methodologies are not clearly described. The absence of such information results in an inability to assess the versatility of these prediction tools based on their development methodologies. It also results in an inability to compare these tools with other prediction tools.

Training and test data sets

In general, all prediction tools discussed in this work are trained using data sets consisting of ≤80 instances. This is in view of the limited number of proteins available with experimentally determined folding rates. This constraint would hamper the performance of the prediction tools by lowering their reliability, and generalization capability [19]. Consequently, the reliability and generalization capability of both K-fold and SeqRate may have been compromised owing to their use of the two smallest training data sets.

Biased prediction outcome and overestimation of performance are foreseeable when there is an extensive overlap between the training and test data sets [19]. SeqRate is the only tool reviewed that has adopted discrete training and independent test data sets. The training and independent test data sets in SeqRate are constructed by random selection of 90 and 10% of the data out of the 61 instances available, respectively. Such a data partitioning strategy has proven to be advantageous because the prediction model is developed on a common training data set while being completely blinded to the test data set [42]. As an initiative to minimize the similarity among subsets during cross-validation, K-Fold has implemented a methodical approach for the partitioning of data by using the BLASTclust program. This program is capable of computing pairwise matches and successively forms clusters with similar sequences. Apart from sequence homology between the training and test data sets, the issue of data redundancy within the data set has to be addressed appropriately. SWFoldRate, in particular, has emphasized on the removal of homologous sequences from the training data set through the use of UniProt sequence comparison (http://www.uniprot.org/) [43]. In view of this, SeqRate, K-Fold and SWFoldRate have training data sets of higher quality compared with other tools that did not use either data partitioning or sequence homology reduction strategies.

Prediction tools that are trained with an unbalanced data set often lead to prediction models with poor performance [36, 44]. In this regard, the types of data imbalance observed in the training data sets of protein folding rate prediction tools can be classified into two different categories, namely, imbalances owing to the structural classification and those owing to the folding kinetics of proteins. The structural classification of a protein is defined by the elements of its secondary structure, such as α-helices and β-strands in the protein. All α-proteins contain >40% α-helices while maintaining <5% of β-strands and vice versa for all β-proteins, while mixed class proteins contain at least 15% α-helices and 10% β-strands [31]. On the other hand, the folding kinetics of a protein are represented by the kinetic order, which indicates the occurrence of an intermediate state during the folding process [11]. Two-state proteins are generally smaller in size and are able to fold at a faster speed [11], whereas multi-state or three-state proteins form intermediates before the formation of the native 3D structure [7]. In brief, the data in SWFoldRate are scaled before SVM application to prevent domination of attributes in greater numeric ranges over those in smaller numeric ranges, as well as to avoid any possible numerical difficulty during the calculation [5]. Similar efforts to minimize the effect of unbalanced data have not been observed in the development of other prediction tools.

Performance evaluation strategy

Among the tools reviewed, a cross-validation test has become popular as the validation and performance gauging approach. Apart from SeqRate, which has adopted an independent test data set, the remaining tools discussed have evaluated the respective predictive ability using cross-validation, predominantly with LOOCV tests. This is in accordance with the noteworthy attribute of LOOCV tests in offering approximately unbiased outcome estimation [45]. In addition, the relatively small sizes of data sets used in these tools actually favor the selection of LOOCV as opposed to other cross-validation methods [42, 45]. In contrast, the independent test conducted by SeqRate allows strict assessment on the generalization capability of SeqRate on unseen data.

SeqRate is the sole tool reviewed that includes experimental verification. Provided that the query sequence was correctly classified in terms of its fold kinetic category, precise estimation of the protein folding rate can be achieved. For instance, the DNA-binding protein Engrailed Homeodomain

(PDB ID: 1ENH), which has an experimentally measured fold rate of 10.5, was predicted to have a fold rate of 10.05 (both values are reported in natural-base logarithm scale) [11]. Experimental verification is an equally important validation method, as this serves as the best evidence with which to assess the applicability and operability of the tool in a real-life scenario. In this regard, SeqRate has adopted a better performance evaluation strategy compared with other tools by validating the prediction model with both an independent test data set and wet laboratory procedures.

Prediction performance

The performance of folding rate prediction tools is mainly evaluated by correlation coefficient and difference measures. The Pearson's correlation coefficient (PCC), also known as the Pearson product-moment correlation coefficient [46], is used to discover the correlation between the experimentally determined and model-predicted values. Various types of difference measures, such as standard error (SE), mean absolute deviation (MAD) and root mean square error (RMSE), are used in various prediction tools to describe the extent of deviation of the model-predicted values relative to the experimentally determined value. Moreover, different validation schemes and varying training data sets are used in all the prediction tools reviewed. Owing to these variations, it is inappropriate to choose the best tool by direct comparison of various indicators, as reported by the respective prediction tool developers.

To make a fair comparison of the performance of various tools, we have conducted an empirical comparison using a standardized non-overlapping test data set to evaluate the performance of each tool. The PCC, RMSE and MAD of each tool were assessed using the standardized non-overlapping test data set, and the results are tabulated in Table 2. This is a non-overlapping test data set independent of the training data sets of each tool. This non-overlapping test data set comprises 28 samples (Supplementary Table S1) that were collected from multiple resources [47–51]. In addition, sequence redundancy reduction has been conducted using the CD-HIT suite [52] at 30% sequence identity to remove highly similar sequences.

When subjected to performance evaluation using the standardized non-overlapping test data set, SeqRate was found to outperform the other tools with the highest PCC and lowest RMSE and MAD. This is mainly owing to the robust prediction tool development methodology, efficient feature selection strategy and high quality of the independent training and test data sets that have been implemented in SeqRate. Pred-PFR achieved the second highest PCC with slightly higher statistical deviations. Surprisingly, negative correlations have been observed for SFoldRate, FOLD-RATE, K-FOLD, PRORATE and SWFoldRate. A notable work published by Willmott has demonstrated that the application of PCC to compare the performances of different models is often inappropriate and misleading. Negative values of PCC are possible even when the model-predicted values do not differ much from the experimentally determined values. In contrast, Willmott has concluded that RMSE and MAD are better overall measures of prediction model performance. Both RMSE and MAD summarize the average difference in the units of the predicted and experimental values, thus giving more helpful and straightforward information regarding the performance of the prediction model [53]. Therefore, in this section, we compare the performance of each model by examining the RMSE and MAD (Table 2). The relatively high RMSE and MAD of both SFoldRate and PRORATE indicate that these two tools performed poorly when assessed objectively using this independent test data set. The low overall performance of these tools might be related to the less robust

Table 2: Empirical comparison of the prediction performance of web-based protein-folding rate prediction tools using the standardized non-overlapping test data set[a]

Tool	PCC	RMSE	MAD
SFoldRate	−0.0603	18.11	11.73
FOLD-RATE	0.3597	4.64	3.85
FOLD-RATE[b]	0.1348	8.84	7.71
Pred-PFR	0.5201	3.86	3.06
FoldRate	0.5038	4.12	3.28
K-Fold	−0.4771	4.56	4.50
PRORATE	−0.1720	94.18	60.79
PRORATE[c]	−0.1720	94.18	60.79
SWFoldRate	−0.3758	5.81	4.39
SeqRate	0.6750	2.46	2.09
SeqRate[c]	0.6349	2.54	2.16

[a]The standardized non-overlapping test data set is available in Supplementary Table S1 (Supplementary Materials). [b]Predictions are conducted with protein structural information as the input. [c]Predictions are conducted with kinetic order information as the input.

training algorithm and the low quality of the training data sets.

The error measures produced from the standardized non-overlapping test data set are called 'out-of-sample' measures. In comparison, error measures that are determined using a training data set along with cross-validation strategies are known as 'in-sample' measures [54]. 'Out-of-sample' error signifies the generalization capability of the prediction model when being applied to unseen data. Therefore, it is of paramount importance to minimize 'out-of-sample' error instead of 'in-sample' error.

From a comparison of the 'out-of-sample' error computed using the standardized non-overlapping test data set and the 'in-sample' error as reported by the respective prediction tool developers, it is apparent that the 'in-sample' errors reported indicate overestimated prediction performance. As mentioned earlier, SeqRate is the only tool that has used an independent test data set for performance assessment. Therefore, the PCC and MAD of SeqRate determined using the standardized non-overlapping test data set agree with the reported values in Table 1 more closely than those of other tools. It is clear that SeqRate performed best out of all the prediction tools assessed in this study.

In addition to conducting prediction without any protein structure or kinetic order information, the performances of FOLD-RATE, PRORATE and SeqRate were re-assessed using the standardized non-overlapping test data set along with the protein structure and kinetic order information. FOLD-RATE adopts a different scheme comprising at least three different prediction models for different protein structural classifications. FOLD-RATE predictions perform more poorly than predictions made without protein structural information. The structural classifications of proteins in this standardized non-overlapping test data set are extracted from the Structural Classification of Proteins (SCOP) database (http://scop.mrc-lmb.cam.ac.uk/scop/). Alternatively, instead of implementing prediction models based on protein structural classification, both SeqRate and PRORATE use individual prediction models for query sequences with two-state and multistate folding kinetics. Despite provision of the kinetic order information, PRORATE was reported to perform similarly to prediction without kinetic order information. SeqRate, on the other hand, recorded slightly lower PCC and slightly greater error measures. Nevertheless, SeqRate remains the most reliable protein-folding rate prediction tool and has an appealing generalization capability when applied to unseen data.

Usability and utility

K-Fold and PRORATE only accept either a Protein Data Bank (PDB) code or a PDB file as input. The PDB is a public archive of structural data for biological macromolecules [55]. The PDB code is the accession code assigned to proteins whose structures are deposited, annotated and validated in (and distributed by) the PDB database. In addition to the accession code, the PDB file is available for download from the database's Web site. Newly discovered protein sequences that have not been assigned a PDB code will not be accessed by either K-Fold or PRORATE. The fact that both K-Fold and PRORATE require the availability of amino acid sequence input in the form of a PDB code or PDB file has restrained the usability of these tools. Another important criterion to judge the usability of a prediction tool is the degree of helpfulness. Pop-up windows, error statements and examples of query sequence, are criteria that can be used to outline the degree of helpfulness. SWFoldRate and SeqRate are unequipped with error statements or pop-up windows even when an error has occurred during execution.

SWFoldRate permits submission of multiple input sequences of up to 10 sequences in FASTA format at one time. Furthermore, batch submission with 50 sequences in FASTA format is available in SWFoldRate. However, the use of batch submission necessitates the upload of a file containing query sequences in FASTA format and the prediction outcome is only available through email. SeqRate also returns the prediction outcome via email instead of direct output displayed on the web page. This reduces the efficiency of the prediction process owing to indirect retrieval of the prediction outcome. In addition to the inconvenience in retrieving the outcome of the prediction, both K-Fold and SeqRate generally require longer computing time compared with the other prediction tools. In addition to returning the protein folding rate, K-Fold, PRORATE and SeqRate can return multiple other outputs following the prediction process. These include structural information on the query sequence, such as contact number [56] and contact order [23].

Table 3: Summary of the respective advantages, disadvantages and preferences of each prediction tool

Prediction tools	SFoldRate	FOLD-RATE	Pred-PFR	FoldRate	K-fold	PRORATE	SWFoldRate	SeqRate
Advantages								
Ability to process ambiguous residue 'X'	Y	–	–	–	Y	Y	Y	Y
Additional output	–	–	–	–	Y	Y	–	Y
Balanced training data set	–	–	–	–	–	–	Y	–
Multiple sequence prediction	–	–	–	–	–	–	Y	–
Sequence redundancy reduction	–	–	–	–	–	–	Y	–
Disadvantages								
Input format constraint	–	–	–	–	Y	Y	–	–
Requirement on additional input	–	Y	–	–	–	Y	–	Y
Training data set <75 data	–	–	–	–	Y	–	–	Y
Restricted applicability to input sequence with ≥50 amino acids	–	–	Y	Y	–	–	–	–
Proposed conditions:								
Multiple input sequences	–	–	–	–	–	–	Preferred	–
Input sequence containing ambiguous residue 'X'	Preferred	–	–	–	Preferred	Preferred	Preferred	Preferred
Input protein with known kinetic order	–	–	–	–	–	Preferred	–	Preferred
Input protein with known structural class	–	Preferred	–	–	–	–	–	–
Input protein with unknown kinetic order and structural class	Preferred	–	Preferred	Preferred	Preferred	–	Preferred	–
Input sequence <50 amino acids	Preferred	Preferred	–	–	Preferred	Preferred	Preferred	Preferred

Y = Yes.

To provide more useful information, a simplified comparison of the eight web-based prediction tools reviewed in this article is presented in Table 3. The respective advantages, disadvantages and preferences of each tool are summarized, along with the guidance on the preferred tool under a variety of possible conditions.

Alternative prediction tools

In addition to the methodologies discussed earlier, an alternative mean of predicting the protein-folding rate is through homologous sequence search via databases such as PPT-DB (http://www.pptdb.ca/), which is essentially a protein property database [50]. However, this method lacks flexibility and reliability because protein-folding rate prediction is unable to be conducted when the input sequence bears little resemblance to those annotated in the database. Moreover, the limited number of proteins available in the database reduces the performance and usability of this approach. PPT-DB can be used to make reliable prediction for approximately 75% of all query sequences provided that this database encompasses >10 000 sequences [50]. Nevertheless, PPT-DB contains only 83 sequences with experimentally determined folding rates, which is far less than the expected amount. As a consequence of this data deficiency, protein-folding rate prediction by similarity searches using this database can be regarded as unfeasible. Apart from PPT-DB, there exist other freely available databases of experimental data regarding protein-folding, including KineticDB [57] and PFD 2.0 [49]. In addition to predicting the protein-folding rate using homologous sequence search, both PFR-AF [47] and the back propagation neural network model [58] are alternative prediction tools for protein-folding rate prediction. However, these two prediction tools were not assessed in this study owing to their inaccessibility through a web-server and their prerequisite of demanding additional inputs before prediction. More specifically, PFR-AF has to be used in combination with secondary structure prediction tools, such as PSI-PRED [59], PROTEUS [60] and SSPRO [61]. The cumbersome procedures involved in PFR-AF are believed to reduce the interest of potential users.

Current limitations and future prospects

In the development of protein-folding rate prediction tools that use the amino acid sequence as the basis of prediction, any sequence-independent factors, such as temperature [62], macromolecular crowding [63] and assistance from molecular chaperones [64], have been disregarded during the

prediction process. Accordingly, the application of these prediction tools is limited and they are incapable of identifying the fluctuations in folding rates that occur under varying experimental conditions. Indisputably, one of the major challenges in improving the existing tools for protein-folding rate prediction is the limited amount of available experimental data on folding rate [11]. Similarly, the relatively small number of proteins with solved structures entails a growing disparity between proteins with known sequences and those with known structures. However, with advanced artificial intelligence and data-mining technology, users of bioinformatics tools should remain optimistic in anticipating an increase in data emergence in the near future. Lastly, being unable to conduct the prediction in the presence of ambiguous residues, the application of the associated prediction tools has been restricted to query sequences that do not contain any ambiguous residues. This limitation applies to FOLD-RATE, Pred-PFR, FoldRate, K-Fold and PRORATE, and reduces the usability of these tools. The existence of ambiguous residues is undoubtedly a great hindrance in efforts to improve the performance of the prediction tools. This is owing to the unknown properties of the ambiguous residues in the amino acid sequence. However, there is a trade-off between the performance and usability of the prediction tool. It is anticipated that future prediction tools will be able to overcome this limitation by accepting the presence of ambiguous residues in the query sequence while being able to achieve a satisfactory prediction performance.

CONCLUSION

The ability to predict protein-folding rates without the need for *in vivo* or *in vitro* experimental work has motivated bioinformaticians to develop bioinformatics tools that can be applied in the field of molecular biology. The comparison of eight distinct web-based tools for protein-folding rate prediction presented in this work can assist researchers in selecting the most appropriate tool in various circumstances. Non-bioinformaticians, in particular, can benefit from this easily comprehensible review that emphasizes the comprehensive comparison of the most widely used web-based protein-folding rate prediction tools. The development method, performance, usability and utility of the prediction tools have been reviewed and compared in depth. Depending on the needs in particular circumstances, different prediction tools reviewed in this work can be used. In general, SeqRate is noted as the best performing tool with the lowest error and highest correlation coefficient, and it can be applied to query sequences with ambiguous residues. With the massive leap in bioinformatics tool development and the accumulation of proteomic data in the past decade, it is worthwhile anticipating that new prediction tools with superior performance and usability will be developed in the near future.

SUPPLEMENTARY DATA

Supplementary data are available online at http://bib.oxfordjournals.org/.

Key Points

- This review serves as a helpful guide in selecting a suitable web-based protein-folding rate prediction tool for use in specific circumstances.
- The performance of each bioinformatics tool has been evaluated based on a standardized non-overlapping test data set.
- The usability and utility of each bioinformatics tool have been critically reviewed.

Acknowledgements

The authors are grateful to Monash University Sunway Campus Malaysia for providing the research support needed for this work. C.C. is a recipient of the Higher Degree by Research Scholarship awarded by Monash University Sunway Campus Malaysia. J.S. is National Health and Medical Research Council of Australia (NHMRC) Peter Doherty Fellow and a recipient of the Hundred Talents Program of CAS and the Monash University Fellowship incubator program.

FUNDING

This work is supported by Ministry of Higher Education (MOHE) of Malaysia [FRGS/1/2012/TK05/MUSM/03/2]; Ministry of Science, Technology and Innovation (MOSTI) of Malaysia [e-Science: 02-02-10-SF0088]; National Natural Science Foundation of China [61202167, 61303169]; Knowledge Innovative Program of the CAS [KSCX2-EW-G-8]; National Health and Medical Research Council of Australia (NHMRC) [490989]; Australian Research Council [LP110200333]; and Major Inter-disciplinary Research (IDR) Project Grant awarded by Monash University and the Hundred Talents Program of CAS.

References

1. Vendruscolo M, Zurdo J, Macphee CE, Dobson CM. Protein folding and misfolding: a paradigm of self-assembly and regulation in complex biological systems. *Philos Trans A Math Phys Eng Sci* 2003;**361**:1205–22.
2. Gianni S, Guydosh NR, Khan F, *et al*. Unifying features in protein-folding mechanisms. *Proc Natl Acad Sci USA* 2003; **100**:13286–91.
3. Hoffmann F, Posten C, Rinas U. Kinetic model of *in vivo* folding and inculsion body formation in recombinant *Escherichia coli*. *Biotechnol Bioeng* 2001;**72**:315–22.
4. Idicula-Thomas S, Kulkarni AJ, Kulkarni BD, *et al*. A support vector machine-based method for predicting the propensity of a protein to be soluble or to form inclusion body on overexpression in *Escherichia coli*. *Bioinformatics* 2006;**22**: 278–84.
5. Cheng X, Xiao X, Wu ZC, *et al*. Swfoldrate: predicting protein folding rates from amino acid sequence with sliding window method. *Proteins* 2013;**81**:140–8.
6. Chou KC, Shen HB. FoldRate: a web-server for predicting protein folding rates from primary sequence. *Open Bioinforma J* 2009;**3**:31–50.
7. Capriotti E, Casadio R. K-Fold: a tool for the prediction of the protein folding kinetic order and rate. *Bioinformatics* 2007;**23**:385–6.
8. Gromiha MM, Huang LT. Machine learning algorithms for predicting protein folding rates and stability of mutant proteins: comparison with statistical methods. *Curr Protein Pept Sci* 2011;**12**:490–502.
9. Song J, Takemoto K, Shen H, *et al*. Prediction of protein folding rates from structural topology and complex network properties. *IPSJ Trans Bioinform* 2010;**3**:40–53.
10. Shen HB, Song JN, Chou KC. Prediction of protein folding rates from primary sequence by fusing multiple sequential features. *J Biomed Sci Eng (JBiSE)* 2009;**2**:136–43.
11. Lin GN, Wang Z, Xu D, Cheng J. SeqRate: sequence-based protein folding type classification and rates prediction. *BMC Bioinformatics* 2010;**11(Suppl 3)**:S1.
12. Ouyang Z, Liang J. Predicting protein folding rates from geometric contact and amino acid sequence. *Protein Sci* 2008;**17**:1256–63.
13. Chang CCH, Song J, Tey BT, Ramanan RN. Bioinformatics approaches for improved recombinant protein production in *Escherichia coli*: protein solubility prediction. *Brief Bioinform* 2014;**15**(6):953–62.
14. Diaz AA, Tomba E, Lennarson R, *et al*. Prediction of protein solubility in *Escherichia coli* using logistic regression. *Biotechnol Bioeng* 2010;**105**:374–83.
15. Xiaohui N, Nana L, Jingbo X, *et al*. Using the concept of Chou's pseudo amino acid composition to predict protein solubility: an approach with entropies in information theory. *J Theor Biol* 2013;**332**:211–17.
16. Huang HL, Charoenkwan P, Kao TF, *et al*. Prediction and analysis of protein solubility using a novel scoring card method with dipeptide composition. *BMC Bioinformatics* 2012;**13(Suppl 17)**:S3.
17. Agostini F, Vendruscolo M, Tartaglia GG. Sequence-based prediction of protein solubility. *J Mol Biol* 2012;**421**: 237–41.
18. Idicula-Thomas S, Balaji PV. Understanding the relationship between the primary structure of proteins and its propensity to be soluble on overexpression in *Escherichia coli*. *Protein Sci* 2005;**14**:582–92.
19. Magnan CN, Randall A, Baldi P. SOLpro: accurate sequence-based prediction of protein solubility. *Bioinformatics* 2009;**25**:2200–07.
20. Smialowski P, Doose G, Torkler P, *et al*. PROSO II—a new method for protein solubility prediction. *FEBS J* 2012;**279**:2192–200.
21. Smialowski P, Martin-Galiano AJ, Mikolajka A, *et al*. Protein solubility: sequence based prediction and experimental verification. *Bioinformatics* 2007;**23**:2536–42.
22. Wilkinson DL, Harrison RG. Predicting the solubility of recombinant proteins in *Escherichia coli*. *Nat Biotechnol* 1991; **9**:443–8.
23. Plaxco KW, Simons KT, Baker D. Contact order, transition state placement and the refolding rates of single domain proteins. *J Mol Biol* 1998;**277**:985–94.
24. Gromiha MM, Selvaraj S. 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* 2001;**310**: 27–32.
25. Zhou H, Zhou Y. Folding rate prediction using total contact distance. *Biophys J* 2002;**82**:458–63.
26. Zhang L, Sun T. Folding rate prediction using n-order contact distance for proteins with two- and three-state folding kinetics. *Biophys Chem* 2005;**113**:9–16.
27. Schwede T, Kopp J, Guex N, Peitsch MC. SWISS-MODEL: an automated protein homology-modeling server. *Nucleic Acids Res* 2003;**31**:3381–5.
28. Socci ND, Onuchic JN. Folding kinetics of protein like heteropolymers. *Protein Folds, A Distance Based Approach* 1996;218–32.
29. Anfinsen CB. Principles that govern the folding of protein chains. *Science* 1973;**181**:223–30.
30. Bucciantini M, Giannoni E, Chiti F, *et al*. Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases. *Nature* 2002;**416**:507–11.
31. Gromiha MM, Thangakani AM, Selvaraj S. FOLD-RATE: prediction of protein folding rates from amino acid sequence. *Nucleic Acids Res* 2006;**34**:W70–4.
32. Breiman L. Statistical modeling: the two cultures. *Stat Sci* 2001;**16**:199–231.
33. Wang D, Keller JM, Andrew Carson C, *et al*. Use of fuzzy-logic-inspired features to improve bacterial recognition through classifier fusion. *IEEE Trans Syst Man Cybern B Cybern* 1998;**28**:583–91.
34. Chou KC, Shen HB. Cell-PLoc: a package of web servers for predicting subcellular localization of proteins in various organisms. *Nat Protoc* 2008;**3**:153–62.
35. Noble WS. What is a support vector machine? *Nat Biotechnol* 2006;**24**:1565–7.
36. Yang ZR. Biological applications of support vector machines. *Brief Bioinform* 2004;**5**:328–38.
37. Chiogna M, Nakhaeizadeh G, Taylor C. Probabilistic symbolic classifiers: an empirical comparison from a statistical perspective. In: *Workshop on Machine Learning and Statistics, Workshop of the Seventh European Conference on Machine Learning, 1994, Italy*. Canada: John Wiley & Sons Canada.
38. Cunningham SJ. Machine learning and statistics: A matter of perspective. Working paper 95/11. Hamilton,

New Zealand: University of Waikato, Department of Computer Science, 1995.

39. Song J, Tan H, Boyd SE, *et al*. Bioinformatic approaches for predicting substrates of proteases. *J Bioinform Comput Biol* 2011;**9**:149–78.
40. Kittler J. Feature set search algorithms. *Pattern Recognit Signal Process* 1978;**41**:60.
41. Larrañaga P, Calvo B, Santana R, *et al*. Machine learning in bioinformatics. *Brief Bioinform* 2006;**7**:86–112.
42. Simon RM, Subramanian J, Li MC, Menezes S. Using cross-validation to evaluate predictive accuracy of survival risk classifiers based on high-dimensional data. *Brief Bioinform* 2011;**12**:203–14.
43. Apweiler R, Bairoch A, Wu CH, *et al*. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* 2004;**32**: D115–19.
44. duVerle DA, Mamitsuka H. A review of statistical methods for prediction of proteolytic cleavage. *Brief Bioinform* 2012; **13**:337–49.
45. Molinaro AM, Simon R, Pfeiffer RM. Prediction error estimation: a comparison of resampling methods. *Bioinformatics* 2005;**21**:3301–7.
46. Rodgers JL, Nicewander WA. Thirteen ways to look at the correlation coefficient. *Am Stat* 1988;**42**:59–66.
47. Gao J, Zhang T, Zhang H, *et al*. Accurate prediction of protein folding rates from sequence and sequence-derived residue flexibility and solvent accessibility. *Proteins* 2010;**78**: 2114–30.
48. De Sancho D, Muñoz V. Integrated prediction of protein folding and unfolding rates from only size and structural class. *Phys Chem Chem Phys* 2011;**13**:17030–43.
49. Fulton KF, Bate MA, Faux NG, *et al*. Protein folding database (PFD 2.0): an online environment for the International Foldeomics Consortium. *Nucleic Acids Res* 2007;**35**:D304–7.
50. Wishart DS, Arndt D, Berjanskii M, *et al*. PPT-DB: the protein property prediction and testing database. *Nucleic Acids Res* 2008;**36**:D222–9.
51. Hagai T, Levy Y. Folding of elongated proteins: conventional or anomalous? *J Am Chem Soc* 2008;**130**:14253–62.
52. Huang Y, Niu B, Gao Y, *et al*. CD-HIT suite: a web server for clustering and comparing biological sequences. *Bioinformatics* 2010;**26**:680–2.
53. Willmott CJ. Some comments on the evaluation of model performance. *Bull Am Meteorol Soc* 1982;**63**:1309–13.
54. Madsen H, Pinson P, Kariniotakis G, *et al*. Standardizing the performance evaluation of short-term wind power prediction models. *Wind Eng* 2005;**29**:475–89.
55. Berman HM, Westbrook J, Feng Z, *et al*. The protein data bank. *Nucleic Acids Res* 2000;**28**:235–42.
56. Yuan Z. Better prediction of protein contact number using a support vector regression analysis of amino acid sequence. *BMC Bioinformatics* 2005;**6**:248.
57. Bogatyreva NS, Osypov AA, Ivankov DN. KineticDB: a database of protein folding kinetics. *Nucleic Acids Res* 2009; **37**:D342–6.
58. Zhang L, Li J, Jiang Z, Xia A. Folding rate prediction based on neural network model. *Polymer* 2003;**44**:1751–6.
59. Bryson K, McGuffin LJ, Marsden RL, *et al*. Protein structure prediction servers at University College London. *Nucleic Acids Res* 2005;**33**:W36–8.
60. Montgomerie S, Sundararaj S, Gallin WJ, Wishart DS. Improving the accuracy of protein secondary structure prediction using structural alignment. *BMC Bioinformatics* 2006; **7**:301.
61. Cheng J, Randall AZ, Sweredoski MJ, Baldi P. SCRATCH: a protein structure and structural feature prediction server. *Nucleic Acids Res* 2005;**33**:W72–6.
62. De Sancho D, Doshi U, Munoz V. Protein folding rates and stability: how much is there beyond size? *J Am Chem Soc* 2009;**131**:2074–5.
63. van den Berg B, Ellis RJ, Dobson CM. Effects of macromolecular crowding on protein folding and aggregation. *EMBO J* 1999;**18**:6927–33.
64. Ellis RJ. Molecular chaperones: assisting assembly in addition to folding. *Trends Biochem Sci* 2006;**31**:395–401.