Published in final edited form as:
Proteins. 2012 January ; 80(1): 81–92. doi:10.1002/prot.23163.

PROTS: A fragment based protein thermo-stability potential

Yunqi Li[1], **Jian Zhang**[2], **David Tai**[1], **C. Russell Middaugh**[3], **Yang Zhang**[2], and **Jianwen Fang**[1,*]

[1]Applied Bioinformatics Laboratory, the University of Kansas, Lawrence, Kansas 66047

[2]Center for Computational Medicine and Bioinformatics, the University of Michigan Medical School, Ann Arbor, Michigan 48109

[3]Department of Pharmaceutical Chemistry, the University of Kansas, Lawrence, Kansas 66047

Abstract

Designing proteins with enhanced thermo-stability has been a main focus of protein engineering because of its theoretical and practical significance. Despite extensive studies in the past years, a general strategy for stabilizing proteins still remains elusive. Thus effective and robust computational algorithms for designing thermo-stable proteins are in critical demand. Here we report PROTS, a sequential and structural four-residue fragment based protein thermo-stability potential. PROTS is derived from a non-redundant representative collection of thousands of thermophilic and mesophilic protein structures and a large set of point mutations with experimentally determined changes of melting temperatures. To the best of our knowledge, PROTS is the first protein stability predictor based on integrated analysis and mining of these two types of data. Besides conventional cross validation and blind testing, we introduce hypothetical reverse mutations as a means of testing the robustness of protein thermo-stability predictors. In all tests, PROTS demonstrates the ability to reliably predict mutation induced thermostability changes as well as classify thermophilic and mesophilic proteins. In addition, this white-box predictor allows easy interpretation of the factors that influence mutation induced protein stability changes at the residue level.

Keywords

protein stability; thermophilic; prediction; datamining; thermostability potential

INTRODUCTION

The ability to design proteins with enhanced thermo-stability is important both theoretically and practically.[1–8] Protein-based drugs have become increasingly attractive because of their high efficiency and low side effects. Unfortunately, many native proteins are only marginally stable under both normal physiological and storage conditions. Drugs based on proteins are often susceptible to physical and chemical degradation that affects their potency and safety during manufacturing, transportation, and storage processes.[9] Therefore enhancing the thermo-stability of a protein drug candidate can be a decisive factor in whether it eventually becomes a marketable pharmaceutical. Enzymes with enhanced stability are also useful in many biotechnological applications. Such enzymes allow

*Correspondence to: Jianwen Fang, Applied Bioinformatics Laboratory, the University of Kansas, 2034 Becker Dr., Lawrence, KS 66047. jwfang@ku.edu.

Additional Supporting Information may be found in the online version of this article.

catalyzed reactions to be performed at higher temperature, which can lead to more efficient industrial processes because chemical reactions are intrinsically faster at higher temperature.[7,8]

Computational methods for designing proteins with enhanced thermostability are attractive due to their potential low cost and time-saving properties over current experimental approaches.[10] In general, these methods attempt to define general principles of protein thermo-stability and apply them to rationally design novel proteins. Despite extensive studies in the past several years[1–3,5]; however, a general strategy for stabilizing proteins remains elusive.[11] This is primarily due to the diverse mechanisms contributing to protein stabilization.[12] Thus effective and robust computational algorithms for designing thermo-stable proteins are still in critical demand.

Thermophiles are organisms which live at elevated temperatures as high as 113°C.[5] Thus, the proteins produced by thermophiles (thermophilic proteins or TPs) are intrinsically more thermo-stable than their mesophilic counterparts (MPs). Consequently one common approach to developing thermo-stable proteins is to perform comparative studies of the sequences and/or structures of TPs and their MPs, in the hope of discovering structural patterns of protein thermo-stabilization.[13–21] For example, Haney *et al.* found an increased level of charged residues in TPs[22] and Glyakina *et al.* found that more closely packing of the external, water-accessible residues.[23] These comparative studies have revealed a number of general trends that produce protein stabilization. It is challenging, however, to identify and apply suitable rules to predict favorable mutations that may enhance the thermo-stability for each individual protein.

Another approach is to use force-fields and potentials, either general purpose ones or those specifically developed for predicting protein stability, to predict mutation induced thermo-stability changes. For example, FoldX provides a quantitative estimation of the contributions of specific interaction to protein stability and has been benchmark-tested on a large set of point mutations.[24] Gu *et al.* developed eScape for analyzing the protein energy lanscape of a protein sequence and showed its correlation with protein stability across proteomes between mesophiles and thermophiles.[25,26] Other notable approaches include LSE,[27] EGAD,[28] DFIRE,[29] and ERIS.[30] ROSSETA, a suite of software programs well-known for its use in protein structure predictions, also has the capacity to make thermostability predictions.[2]

In recent years, data mining technologies employing various machine learning algorithms have increasingly attracted attention. Algorithms such as support vector machines,[31–34] neuronal networks,[35] or multiple regression and classification techniques,[36,37] have been used for predicting protein stability changes induced by mutations. The general procedure of machine learning approaches is to train predictive models based on available experimental data using features (properties) such as substitution types, secondary structure, solvent accessibilities, and the presence of neighboring residues. These approaches hold great promises because they may be used to discover subtle patterns governing mutation induced stability changes and protein stability in general. The drawback associated with these types of approaches is also obvious because these models were trained and tested on mutations from a relatively small set of proteins due to the lack of availability of experimental data at the time of their construction.[30] For example, Cheng *et al.* developed a support vector machine predictive model based on 1023 mutations in 36 proteins.[32] This number is rather small if one considers the fact that there are 380 different types of single mutations. As Dokholyan *et al.* pointed out, “The improvement of the prediction accuracy relies on the available experimental stability data for parameter trainings. It is questionable whether parameters obtained from these trainings are transferable to other protein studies”.[30] Thus the robustness of these methods needs to be further validated on larger datasets.

Here we report PROTS, a novel sequential and spatial fragment based PROtein Thermo-Stability potential, which integrates TP/MP comparative analysis and experimental mutation data mining. We create a comprehensive and non-redundant set of high-resolution protein structures of TPs and MPs. Fragments consisting of four amino acid residues were chosen as the atomic units for determining the overall thermo-stability of proteins. The frequencies of sequential tetrapeptides and spatial Delaunay tetrahedrons (DT)[38] in TPs, MPs, and protein mutants are analyzed, and a lookup table is created for calculating the PROTS potentials of proteins and their mutants. We suggest that these two types of data can be integrated because HP/MP orthologs are essentially equivalent to mutants of each other.

Structural information can generally improve the performance of protein property prediction algorithms. The vast majority of proteins, however, lack solved structures. Fortunately, current state-of-the-art protein homologous modeling algorithms are able to produce practically useful structural models.[39] In this work, we test the PROTS potential in homolog models, created using the I-TASSER algorithm,[40–42] of 540 pairs of TP/MP orthologs.[43]

In this work, we introduce hypothetical reversed mutations to test the robustness of computational methods for predicting protein stability changes upon mutations. Usually protein stability changes upon mutations are experimentally measured through changes in the melting temperature (ΔT_m) or alteration of folding free energies ($\Delta\Delta G$) between a wild type protein and its mutant. Existing protein stability predictors use one or the other as the metric for stability changes. Both metrics are thermodynamic parameters and thus state functions.[44] Therefore, the ΔT_m of a mutation from a wild type protein to its mutant ($\Delta T_{mWt\rightarrow Mu}$) equals the negated ΔT_m of a hypothetical reversed mutation (from the mutant to the wild type protein, $\Delta T_{mMu\rightarrow Wt}$):

$$\Delta T_{mWt\rightarrow Mu} = -\Delta T_{mMu\rightarrow Wt} \quad (1)$$

$$\Delta\Delta G_{Wt\rightarrow Mu} = -\Delta\Delta G_{Mu\rightarrow Wt} \quad (2)$$

A robust predictor should treat ΔT_m and $\Delta\Delta\Delta G$ as thermodynamic parameters and be able to achieve identical or at least similar performance on hypothetical reversed mutations to the forward mutations. Our study described below indicates that these tested machine learning algorithms are not robust in such a test.

In the following sections, we describe the applications of the potential to predicting stability change upon mutations, as well as discriminating MP/TP native structures and homolog models. We will also present a comparison of PROTS to several other relevant potentials or algorithms, in the classification of thermophilic/mesophilic proteins and the prediction of protein stability changes upon mutations. In all cases, PROTS compares favorably. We describe the procedure of collecting training and test datasets, and then the construction of the lookup table used for computing the PROTS potential in the Experimental Procedures.

MATERIALS AND METHODS

Nonredundant TP/MP native structures

In this study, we use a collection of nonredundant 1020 TP and 4742 MP structures that was previous used in developing distance-dependent statistical potentials for discriminating TPs and MPs and the procedure was described previously.[45] Table I in Text S1 provides a complete list of the organisms and distribution of these proteins in each organism.

Structural modeling of 540 TP/MP ortholog pairs

Structural models of 540 TP/MP ortholog pairs, which did not have structures in the PDB library, are predicted using I-TASSER.[40–42] These ortholog pairs were previously used in sequence-based TP/MP classification and relative thermostability prediction.[43] I-TASSER is a hierarchical approach to both template-based and *ab initio* modeling of protein structures, and it was ranked as the best methods for automated protein structure prediction in communitywide blind experiments, CASP7 and CASP8.[46,47] For a given target sequence, I-TASSER first identifies template structure and sequence-structure alignments by LOMETS, a locally installed meta-threading algorithm including 9 start-of-the-art threading programs.[48] Continuous fragments of length >5 residues are then used to reassemble the global topology of a protein under the guide of consensus restraints from multiple threading templates. The structural assembly is performed by replica exchange Monte Carlo simulations. The simulation trajectory decoys are then clustered to identify lowest free energy C_α-represented models using SPICKER.[49] Finally, all-atom models are constructed based on the reduced C_α model using REMO through optimizing the hydrogen-bonding network.[50]

The accuracy of the I-TASSER models can be reliably estimated by the confidence score (C-score) which is a combination of the Z-score of the threading templates in LOMETS and the structure density of SPICKER. In a recent large benchmark study,[51] it was shown that the Pearson correlation coefficient of C-score and the TM-score (a measure of structural similarity to the native structure[52]) is 0.91. For these 540 TP/MP ortholog pairs, there are 97% of cases where the C-score is higher than −1.5, a cutoff for I-TASSER models of correct topology; there are 99% of cases where there is at least one threading template which has the Z-score higher than the inherent Z-score cutoff (meaning the template is a significant hit in threading). Thus, the majority of I-TASSER models are anticipated to have correct topology, which guarantees the quality of corresponding structure-based analyses.

Mutation datasets

We collect a set of point mutations with known melting temperatures (T_m) from the Protherm database.53 Mutations with absolute ΔT_m less than 1°C are excluded because such small changes may not be statistically significant.[54] For mutations with multiple ΔT_m values, we use the median ΔT_m of these mutations if the sign of all ΔT_m values is consistent and excluded them otherwise. The final dataset includes 1146 mutants from 100 different wild type proteins. These proteins are clustered using BLASTClust[55] with a sequence identity threshold of 30%. We obtain 84 distinct clusters and then split them into five groups, each with approximately the same number of mutations, for cross validation. In the cross validation test, mutations in four out of five groups are used for training and the mutations in the remaining group are used for testing. This procedure is repeated four more times until every mutation is used once.

We also obtained a set of point mutations with known free energy changes ($\Delta\Delta G$) from the literature for testing purposes.[11] This dataset contains 2156 single-point mutations from 84 wild type proteins and was previously used in a comparative study of different approaches to predict mutation induced stability changes.[11]

In addition, a set of wild type proteins and their mutants, all with known structures, were collected from the Protherm database. We only consider structure pairs with known $\Delta\Delta G$ of the mutations with resolution of protein structures better than 2.2 Å. There are 155 structure pairs, including 140 for single mutations, originated from nine different wild type proteins in the dataset (Table II in Text S1).

Hypothetical reversed mutations as testing datasets

Currently available mutation induced stability change data, especially those available in the Protherm database,[53] have been widely used in protein stability prediction algorithm development. Therefore using this data to test existing algorithms may not provide an accurate test of performance because of the potential overfitting problem. In this study we adopt a novel approach to construct testing datasets by using hypothetical reversed mutations based on the fact that the melting temperature and free energy are thermodynamic state functions [Eqs. (1,2)].

Secondary structure and solvent accessibility assignment

We use DSSP[56] to assign the secondary structure states and solvent accessible status of all residues in proteins. Each residue is assigned to one of the three classes of secondary structure (helix/strand/coil). We use three levels of solvent accessibility: buried, intermediate, and exposed residues. The solvent accessible area ratio (normalized by the maximum solvent accessible area of each amino acid) of a buried residue is less than 0.25 and an exposed residue is larger than 0.5. All others are assigned as intermediate residues.

PROTS

Two types of four-residue fragments in proteins are used to calculate the PROTS potential. The first type includes all 20^4 sequential tetrapeptides (abbreviated as SEQ), the full permutation of four amino acids. The other comprises the 8855 spatial DTs,[38] the exhaustive combination of four amino acids.

All DTs are grouped into three categories according to the number of the continuously sequential residues in the DTs. Type D43 contains the DTs formed by at least three continuous residues. Type D2 contains at least one two-continuous-residues motif but not extending to three continuous residues. Type D1 is formed by four non-neighboring residues.[38,57,58] We only include the DTs with maximal edge equal or less than 12 Å.[59]

Since the structures of mutants are usually unavailable, we assume that point mutations do not cause significant conformational changes and therefore the structures of mutants are created by simply replacing the wild type residues with mutated residues.

Each sequential fragment in PROTS has 13 features and each spatial fragment has 7 DT features. The 13 sequential features include seven potential terms [calculated by Eq. (6)] including dS(occurrence, W_i), dS(helix, W_i), dS(strand, W_i), dS(coil, W_i), dS(expose, W_i), dS(bury, W_i), dS(intermediate, W_i), and six propensity terms including dD(helix, W_i), dD(strand, W_i), dD(coil, W_i), dD(expose, W_i), dD(bury, W_i), and dD(intermediate, W_i). The 7 DT features include dS(occurrence_DT, W_i), dS(D43, W_i), dS(D2, W_i), dS(D1, W_i) and the propensity terms dD(D43, W_i), dD(D2, W_i) and dD(D1, W_i).

The occurrence probability of a given structural feature K (e.g. helix, strand, coil) for a fragment W_i in a given training dataset X, $P_X(K, W_i)$, is calculated using Eq. (3):

$$P_X(K, W_i) = \frac{N_X(K, W_i)}{\sum_i N_X(K, W_i)} \quad (3)$$

Here i runs over all possible four-residue fragments and $N_X(K, W_i)$ is the number of fragments W_i for a feature K in a given dataset X. P_X(occurrence, W_i) is the occurrence probability of the fragment W_i in the dataset X. The propensity for the W_i in the structure state indicated by feature K is defined as

$$D_{\mathrm{X}}(\mathrm{K}, \mathrm{W}_i)=\frac{P_{\mathrm{X}}(\mathrm{K}, \mathrm{W}_i)}{P_{\mathrm{X}}(\mathrm{occurrence}, \mathrm{W}_i)} \quad (4)$$

We also calculate the Shannon entropy of all fragments defined as

$$S_{\mathrm{X}}(\mathrm{K}, \mathrm{W}_i)=-P_{\mathrm{X}}(\mathrm{K}, \mathrm{W}_i)\ln P_{\mathrm{X}}(\mathrm{K}, \mathrm{W}_i) \quad (5)$$

The potential contribution of feature K of a fragment $\mathrm{W_i}$, $dS(\mathrm{K}, \mathrm{W_i})$ is defined as:

$$dS(\mathrm{K}, \mathrm{W_i})=S_{\mathrm{T}}(\mathrm{K}, \mathrm{W_i})-S_{\mathrm{M}}(\mathrm{K}, \mathrm{W_i}) \quad (6)$$

Here T and M are the sets of TPs and MPs, respectively. Using Eq. (6), we calculate the potential contributions of all features of all fragments from native protein structures. Similarly, we can calculate the propensity difference $dD(\mathrm{K}, \mathrm{W_i})$. The Shannon entropy is not used for propensities because they distribute over a small number of structural features while the four-residue fragments are distributed over a large number of types ($>10^3$).

TP and MP orthologs are essentially mutants with multiple mutations of each other. Thus in principle TP/MP and mutation data are equivalent. We classify all fragments involved in mutations into stabilizing or destabilizing fragments according to the thermo-stability changes caused by the mutations. The stabilizing (ST) fragments are those found in mutants in stabilizing mutations or from wild type proteins in destabilizing mutations. The destabilizing (DE) fragments are from mutants in destabilizing mutations or from wild type proteins in stabilizing mutations. The Eq. (6) is revised to

$$dS(\mathrm{K}, \mathrm{W}_i)=S_{\mathrm{T}}(\mathrm{K}, \mathrm{W}_i)-S_{\mathrm{M}}(\mathrm{K}, \mathrm{W}_i)+\delta_{\mathrm{ST}}(\mathrm{W}_i)S_{\mathrm{T}}(\mathrm{K})-\delta_{\mathrm{DE}}(\mathrm{W}_i)S_{\mathrm{M}}(\mathrm{K}) \quad (7)$$

Here the first two terms are derived from native TP and MP structures and the last two are calculated from the point mutation dataset. $S_{\mathrm{T}}(\mathrm{K})$ and $S_{\mathrm{M}}(\mathrm{K})$ are the potential terms corresponding to the most popular four-residue fragments from TPs and MPs, respectively. The factors $\delta_{\mathrm{ST}}(\mathrm{W_i})$ and $\delta_{\mathrm{DE}}(\mathrm{W_i})$ are used to address the thermo-stability preference of fragments based on the point mutation dataset:

$$\delta_{\mathrm{ST}}(\mathrm{W_i})=\frac{n_{\mathrm{ST,Mu}}(\mathrm{W_i})+n_{\mathrm{DE,Wt}}(\mathrm{W_i})}{\sum n(\mathrm{W_i})} \quad \text{and} \quad \delta_{\mathrm{DE}}(\mathrm{W_i})=\frac{n_{\mathrm{ST,Wt}}(\mathrm{W_i})+n_{\mathrm{DE,Mu}}(\mathrm{W_i})}{\sum n(\mathrm{W_i})} \quad (8)$$

Here, the denominator is the total number of occurrences of a given fragment in the training dataset, Wt and Mu represent wild type proteins and mutants, respectively.

The thermo-stability potential P for a given protein is calculated using:

$$P=-\frac{1}{L}\left\{\sum_{\mathrm{i}}\sum_{\mathrm{K}}\alpha_{\mathrm{K}}dS(\mathrm{K}, \mathrm{W_i})+\sum_{\mathrm{i}}\sum_{\mathrm{K}}\beta_{\mathrm{K}}dD(\mathrm{K}, \mathrm{W_i})\right\} \quad (9)$$

where L is the number of residues in the protein, i runs over all possible sequential and DT spatial fragments, and K includes all 13 sequential and/or 7 DT features.

Since the stability change equals the relative stability difference between mutants and their wild type proteins, the PROTS potential change of a mutation can be calculated by

$$dP = P_{Mu} - P_{Wt} \tag{10}$$

The weights α_K and β_K, the relative contributions of various terms, for the PROTS potential are optimized through maximizing the Pearson correlation coefficient between the predicted stability change ΔP and the experimental observed ΔT_m values based on mutations in the training set. The correlation coefficient R is defined as

$$R = \frac{\sum(\Delta S - \langle \Delta S \rangle)(\Delta T_m - \langle \Delta T_m \rangle)}{sqrt\{\mathrm{Var}(\Delta S)\mathrm{Var}(\Delta T_m)\}} \tag{11}$$

where the numerator is a summation over all mutations in the training dataset, $\langle\ \rangle$ and Var( ) are the mean values and the variance of the variable enclosed.

The PROTS potential can be used to predict thermostability changes whether the protein structure is available or not. All 20 features are used for proteins with structures while only 13 sequential features are used without structures (PROTS_SEQ).

Algorithms used for comparison

To evaluate the performance of PROTS, we compare it to several existing state-of-the-art algorithms for predicting mutation induced thermo-stability changes. FoldX (version 3.0 beta3) is a quantitative estimate of the contributions of interactions to protein stability with a benchmark test on a large set of point mutations.[24] LSE is a statistical local structure entropy derived from representative protein domains, which has demonstrated strong correlation with protein thermostability.[27] MUpro is a support vector machine (SVM) based predictor at sequence level for the variation of folding free energy ($\Delta\Delta G$) upon point mutations.[32] I-Mutant2.0 is a SVM based predictor using structure and sequence information for $\Delta\Delta G$ prediction.[31] EGAD is a force filed based empirical approach to calculate protein stability with rotamer swapping on a fixed backbone scaffold which was shown reliable predictions for more than 1500 mutations.[28]

Performance metrics

The discrimination of thermophilic/mesophilic proteins and stabilized/destabilized mutations can be regarded as a binary classification problem. We generate the receiver operating characteristic (ROC) curve according to the predicted potentials for TPs and MPs, or the potential difference between wild type proteins and their mutants. ROC is a plot of the true-positive ratio (sensitivity) against the false-positive ratio (1−specificity). The area under an ROC curve (AUC) represents the trade-off between sensitivity and specificity. The accuracy of the classification defined as

$$\mathrm{ACC} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \tag{12}$$

We calculate the accuracy at a fixed specificity of 0.80 so that we can directly compare the accuracies of the different models. In this equation, TP, TN, FP, FN stand for true positive, true negative, false positive, and false negative, respectively. A true case represents the class of a protein has been correctly identified. A positive case represents the class of TPs or stabilizing mutations.

We also perform regression analysis of predicted PROTS changes against the ΔT_m or $\Delta\Delta G$ of mutations. The standard regression coefficient R defined in Eq. (11) is used as a metric of the regression performance.

RESULTS AND DISCUSSION

In this section, we first describe parameterizing the PROTS potential based on standard fivefold cross validation in the 1146 point mutations with ΔT_m measurements. This potential is then tested in discriminating native TP/MP structures. We also compare the prediction performance over a large set of point mutations with other algorithms.

Cross validation

We use a standard fivefold cross validation to optimize the weights of all terms in Eq. (9). The absolute values of all weights are restricted to the range of 0–1. We randomly assign an initial weight to each of α_K and β_K in Eq. (9) and then calculate the correlation coefficient R-value. The weights are then randomly updated and the R-value is recalculated. The new weights are kept only if the R-value increased; otherwise the weights are rolled back to the previous values. This procedure is repeated until the R-value reaches a stable plateau. The optimization procedure of the R-values is illustrated in Figure 1 in Text S1. After 5×10^6 steps of optimization, the correlation coefficient reaches 0.653 ± 0.020 in the fivefold cross validation. The quite small error indicates the performance of all classifiers is consistent.

Using the optimized weights in each training fold, we calculate the potentials of mutations in the corresponding holdout testing set for classification and regression analysis. We then calculate the regression R-value of the predicted values against experimentally observed ΔT_m values. The binary classification analysis is performed using $\Delta T_m = 0$ as the threshold to classify mutations as stabilizing or destabilizing. In addition, we use other algorithms to predict ΔT_m of all 1146 point mutations and then perform the same regression and classification performance analysis. Both the regression and the classification results are plotted in Figure 1 and summarized in Table I. PROTS clearly results in favorable classification performance over the other algorithms. For the regression, PROTS also achieves higher correlation coefficients than other methods after mutations used as training data are removed.

The final optimized weights from all five folds are quite similar. We therefore build the final PROTS function by using the averaged weights from the cross validation test and use this in the blind tests presented in the following sections.

PROTS for predicting ΔΔ*G* of single-point mutations

Unlike PROTS, most of other existing algorithms for prediction of mutation induced stability changes were trained and tested on mutations with $\Delta\Delta G$ measurements. We compare the performance of the PROTS potential with other algorithms based on a large set of point mutations with $\Delta\Delta G$ values in both regression and classification analysis. For a fair comparison, mutations used in the training dataset of each algorithm are excluded. The results are presented in Figure 2 and Table II. Clearly PROTS performs better than the other algorithms in the classification of $\Delta\Delta G$ data even though PROTS is developed using TP/MP and ΔT_m data while the others are based on $\Delta\Delta G$ data.

Using hypothetical reverse mutations as a testing dataset

As discussed earlier, both melting temperature and free energy are state functions and therefore the ΔT_m and $\Delta\Delta G$ of a mutation and its hypothetical reverse mutation should obey Eqs. (1) and (2). PROTS performs equally well for the reverse mutations. FoldX and

EGAD, both empirical force field-based predictors, are expected to deliver very similar results. However, the prediction power of machine learning based approaches, that is, MUpro and I-Mutant2.0, diminishes with the hypothetical reversed mutations since their AUCs are close to 0.5 (Tables I and II). LSE is a state function and thus its performance in predicting hypothetical reverse mutations is identical to the forward ones. Its performance is, however, not impressive in either direction (AUC = 0.577, $R = 0.155$). It should be pointed out that for the structure-based predictions made by I-Mutant2.0 and PROTS, the wild type protein structures in the hypothetical reversed mutations are generated by simple substitution of wild type residues with mutant ones without any conformation optimization.

Mutations may alter protein conformations. Therefore, a simple residue substitution without conformation optimization may not reflect reality. To perform a more strict evaluation of the prediction of the hypothetical reversed mutations, we make and evaluate predictions of $\Delta\Delta G$ of 155 mutations with known 3D structures for both wild type and mutants (Table III). Similar to the above test, both MUpro and I-Mutant2.0 deliver significantly different performance for the forward and hypothetical reverse mutations. We use either wild type or mutant structures, respectively, for forward and reverse mutations while using the I-Mutant2.0 (Table III).

The prediction performance of PROTS on reversed mutations is only slightly different from forward ones because the DT features are not identical whether the structure of the wild type protein or its mutant is used. Therefore we test using both structures in the predictions (Table III). As expected, the performance using both structures is slightly better than using only one of them ($R = 0.521$ vs. $R = 0.455$ or 0.447; AUC = 0.862 vs. AUC = 0.844 or 0.838). Such an approach, however, is not very practically useful because the structures of mutants are often unavailable due to the current absence of some structures. Our results, nevertheless, confirm that the current single structure approach assuming no significant conformation changes caused by single mutation is acceptable for stability prediction purposes.

PROTS for discriminating TPs and MPs

Using the optimized weights, we calculate PROTS values for all 1020 TPs and 4977 MPs according to Eq. (9). The ROC curve of the classification is plotted and displayed in Figure 3. In addition to PROTS using all features, we also calculate the values using 13 SEQ or 7 DT features. The AUC of these three functions (PROTS, PROTS_SEQ, and PROTS_DT) are 0.936, 0.903 and 0.889, and the accuracies are 91, 84, and 82%, respectively. Therefore the model using both sequence and DT features achieves better performance than models using either subset of the features. It is clear that both spatial and sequential features are useful for discriminating TPs and MPs.

PROTS shows comparable or better performance on TP/MP classification in comparison to other approaches. For example, Gromiha *et al.* obtained an accuracy of 89% in discrimination of 1609 thermophilic proteins from 3075 mesophilic proteins based on neural network analysis in a fivefold cross validation.[60] TargetStar, a scoring function based on the analysis of 1006 decoy structures for a given protein, can discriminate HP/MP orthologs pairs with 77% accuracy.[61] More recently, Montanucci *et al.* reported a SVM model which achieves 88% accuracy on a set of redundancy-reduced HP/MP pairs.[34]

PROTS for classifying structural models of TPs and MPs

We evaluate the performance of PROTS on classifying TP and MP structure models. We group these proteins into two categories using 30% maximum sequence identity against all of the protein in the training dataset as the cutting threshold. We calculate the PROTS

potentials of the models of all ortholog pairs in these two categories using PROTS and PROTS_SEQ algorithms (Table IV). The accuracies of the pair-wise comparisons of TP/MP orthologs in both categories (94.2% and 97.2%) using PROTS are higher than those using the PROTS_SEQ potential (91.3% and 93.8%), suggesting the structure models built using i-TASSER are useful for such an application. In addition, the difference in accuracies between the close and the distant pairs is fairly small, strongly indicating that PROTS is a robust classifier for discriminating thermophilic/mesophilic protein pairs.

Evaluating the applicability of PROTS

For predicting mutation induced stability changes, it is highly desirable to develop algorithms applicable to many different types of proteins. We define applicability as the ratio of proteins with positive correlation over all proteins in the study because an algorithm can only be applicable to proteins with positive correlation. To evaluate the applicability of PROTS, we select proteins with two or more mutations in ΔT_m or $\Delta\Delta G$ datasets and calculate the correlation coefficients of predicted ΔT_m and $\Delta\Delta G$ versus experimental data for the mutations of each protein (Table V). Using the applicability as metric, PROTS outperforms other approaches in the prediction of mutation induced stability change in the ΔT_m dataset and is among the best in $\Delta\Delta G$ predictions. In both cases, the applicability of PROTS and PROTS_SEQ is higher than 80%. Therefore these algorithms are practically useful in real-world applications.

Analysis of PROTS predictions

We analyze the mutants of three proteins with typical structures: alpha, alpha/beta and beta (Fig. 4). In the prediction of 27 mutants with ΔT_m ranging from −13.1°C to 4.7°C from an alpha-protein (PDB ID: 4LYZ, *Gallus Gallus*), a correlation coefficient of 0.714 is achieved between PROTS predicted stability changes and observed ΔT_m values. Similarly, a correlation coefficient of 0.877 is obtained based on nine mutants from a beta-protein (PDB ID: 2AFG, *Homo Sapiens*) while a correlation coefficient of 0.721 is obtained based on 16 mutants from an alpha/beta protein (PDB ID: 3SSI, *Streptomyces Albogriseolus*). Thus the predicted stability changes at the residue level show strong correlation with experimentally measured ΔT_m values.

All predictive models fall into three categories according to their comprehensibility: white-, gray-, or black-box approaches. The process of a white-box approach is very transparent and well understood by the user. The black-box approach does not allow explicit explanation of the model and the gray-box approaches are partially visible and reasonably understood by the user. PROTS is a white-box approach since the weights of the features determining whether the mutation would stabilize the target protein are known. This may reveal the mechanisms of thermal stabilization. For example, the stabilizing mutation 2AFG-H93G can be largely attributed to the positive contribution from the potential and the propensity from strand/coil and exposure, matching the status of the H93 residue in a surface turn.[62] The destabilizing mutation 2AFG-C83S is caused by the unfavorable changes of the potential and the propensity of coil and exposure, as well as D1 Delaunay tetrahedrons, which agrees with the fact that this residue is located in a core region.[63] The values of all features of these two mutations are listed in Table VI.

ADDITIONAL DISCUSSION

PROTS has two versions. One uses structural information, in addition to sequence information, for target proteins with solved 3D structures. The other version uses only sequence information. Although the sequence-only model is not as accurate as the other that uses both structural and sequential information, the sequence only model, still delivers

reasonably good performance. Such flexibility presents an advantage over force-fields and energy functions, which require high resolution protein structures. Although some machine learning based algorithms can predict protein thermostability based on protein sequences only, these algorithms as we show in this study fail to make acceptable predictions for hypothetical reverse mutations. Therefore, further validation is necessary to establish their robustness.

In this study, we use sequential and spatial fragments consisting of four amino acid residues as the atomic units for determining the overall thermo-stability of proteins. Although it is conceivable that using a larger size of protein fragments may improve the quality and predictive ability of the relative potential, four-residue fragments are practically the largest context for protein sequence and structure data mining because of the limited number of available structures.[64] There are 20^4 different permutations and 8855 combinations of four deposited in the protein data bank (PDB) and the latter amino acid residues. The former number is of the same is close to the number of currently known structural magnitude of protein sequences with solved structures domains.[65] There have been several successful studies using four-residue fragments as the context of protein properties. For example, Chan *et al.* developed tetrapeptide-based local structure entropy,[27] which was later utilized by Bae *et al.* [71] to design and eventually produce stabilized *adenylate kinase* mutants.[66] Using a scoring function based on four-residue Delaunay Tetrahedrons (DTs), Deutsch and Krishnamoorthy were able to discriminate the stability and reactivity changes resulting from mutations with high accuracy.[59]

CONCLUSION

In this work, we develop PROTS, a sequential and spatial fragment based potential, for classifying TPs/MPs and stability changes upon mutations. Our approach utilizes structural profile enhanced lookup tables and exhibits good performance in both classification and regression. We also introduce hypothetical reversed mutations for comprehensive evaluation of the algorithms for protein thermo-stability change predictions. Currently we are applying PROTS to the design of stable mutants of several proteins. The results will be reported separately at a later date.

Acknowledgments

The authors wish to thank the two anonymous reviewers and the editor for their constructive comments and suggestions. They are indebted to Dr. Vladimir Potapov for kindly sharing his data with us and the authors of FoldX, MUpro, I-mutant2.0 and LSE for making their programs and data available.

References

1. Dahiyat BI. In silico design for protein stabilization. Curr Opin Biotech. 1999; 10:387–390. [PubMed: 10449321]
2. Korkegian A, Black ME, Baker D, Stoddard BL. Computational thermostabilization of an enzyme. Science. 2005; 308:857–860. [PubMed: 15879217]
3. Lazar GA, Marshall SA, Plecs JJ, Mayo SL, Desjarlais JR. Designing proteins for therapeutic applications. Curr Opin Struct Biol. 2003; 13:513–518. [PubMed: 12948782]
4. Schweiker KL, Makhatadze GI. Protein stabilization by the rational design of surface charge-charge interactions. Methods Mol Biology. 2009; 490:261–283.
5. Sterner R, Liebl W. Thermophilic adaptation of proteins. Crit Rev Biochem Mol Biol. 2001; 36:39–106. [PubMed: 11256505]
6. Chennamsetty N, Voynov V, Kayser V, Helk B, Trout BL. Design of therapeutic proteins with enhanced stability. Proc Natl Acad Sci USA. 2009; 106:11937–11942. [PubMed: 19571001]

7. Unsworth LD, van der Oost J, Koutsopoulos S. Hyperthermophilic enzymes—stability, activity and implementation strategies for high temperature applications. FEBS J. 2007; 274:4044–4056. [PubMed: 17683334]
8. Schoemaker HE, Mink D, Wubbolts MG. Dispelling the myths—biocatalysis in industrial synthesis. Science. 2003; 299:1694–1697. [PubMed: 12637735]
9. Frokjaer S, Otzen DE. Protein drug stability: a formulation challenge. Nat Rev Drug Discov. 2005; 4:298–306. [PubMed: 15803194]
10. Lippow SM, Tidor B. Progress in computational protein design. Curr Opin Biotech. 2007; 18:305–311. [PubMed: 17644370]
11. Potapov V, Cohen M, Schreiber G. Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details. Protein Eng Des Sel. 2009; 22:553–560. [PubMed: 19561092]
12. Berezovsky IN, Shakhnovich EI. Physics and evolution of thermophilic adaptation. Proc Natl Acad Sci USA. 2005; 102:12742–12747. [PubMed: 16120678]
13. Berezovsky IN, Zeldovich KB, Shakhnovich EI. Positive and negative design in stability and thermal adaptation of natural proteins. PLoS Comput Biol. 2007; 3:e52. [PubMed: 17381236]
14. Gianese G, Argos P, Pascarella S. Structural adaptation of enzymes to low temperatures. Protein Eng. 2001; 14:141–148. [PubMed: 11342709]
15. Mandrich L, Pezzullo M, Del Vecchio P, Barone G, Rossi M, Manco G. Analysis of thermal adaptation in the HSL enzyme family. J Mol Biol. 2004; 335:357–369. [PubMed: 14659763]
16. McDonald JH. Patterns of temperature adaptation in proteins from the bacteria Deinococcus radiodurans and Thermus thermophilus. Mol Biol Evol. 2001; 18:741–749. [PubMed: 11319258]
17. Menendez-Arias L, Argos P. Engineering protein thermal stability. Sequence statistics point to residue substitutions in alpha-helices. J Mol Biol. 1989; 206:397–406. [PubMed: 2716053]
18. Metpally RP, Reddy BV. Comparative proteome analysis of psychrophilic versus mesophilic bacterial species: insights into the molecular basis of cold adaptation of proteins. BMC Genomics. 2009; 10:11. [PubMed: 19133128]
19. Razvi A, Scholtz JM. Lessons in stability from thermophilic proteins. Protein Sci. 2006; 15:1569–1578. [PubMed: 16815912]
20. Zeldovich KB, Berezovsky IN, Shakhnovich EI. Protein and DNA sequence determinants of thermophilic adaptation. PLoS Comput Biol. 2007; 3:e5. [PubMed: 17222055]
21. Zhou XX, Wang YB, Pan YJ, Li WF. Differences in amino acids composition and coupling patterns between mesophilic and thermophilic proteins. Amino Acids. 2008; 34:25–33. [PubMed: 17710363]
22. Haney PJ, Badger JH, Buldak GL, Reich CI, Woese CR, Olsen GJ. Thermal adaptation analyzed by comparison of protein sequences from mesophilic and extremely thermophilic Methanococcus species. Proc Natl Acad Sci USA. 1999; 96:3578–3583. [PubMed: 10097079]
23. Glyakina AV, Garbuzynskiy SO, Lobanov MY, Galzitskaya OV. Different packing of external residues can explain differences in the thermostability of proteins from thermophilic and mesophilic organisms. Bioinformatics. 2007; 23:2231–2238. [PubMed: 17599925]
24. Guerois R, Nielsen JE, Serrano L. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. J Mol Biol. 2002; 320:369–387. [PubMed: 12079393]
25. Gu J, Hilser VJ. Sequence-based analysis of protein energy landscapes reveals nonuniform thermal adaptation within the proteome. Mol Biol Evol. 2009; 26:2217–2227. [PubMed: 19592668]
26. Gu J, Hilser VJ. Predicting the energetics of conformational fluctuations in proteins from sequence: a strategy for profiling the proteome. Structure. 2008; 16:1627–1637. [PubMed: 19000815]
27. Chan CH, Liang HK, Hsiao NW, Ko MT, Lyu PC, Hwang JK. Relationship between local structural entropy and protein thermostability. Proteins. 2004; 57:684–691. [PubMed: 15532068]
28. Pokala N, Handel TM. Energy functions for protein design: adjustment with protein-protein complex affinities, models for the unfolded state, and negative design of solubility and specificity. J Mol Biol. 2005; 347:203–227. [PubMed: 15733929]

29. Zhou H, Zhou Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci. 2002; 11:2714–2726. [PubMed: 12381853]
30. Yin S, Ding F, Dokholyan NV. Modeling backbone flexibility improves protein stability estimation. Structure. 2007; 15:1567–1576. [PubMed: 18073107]
31. Capriotti E, Fariselli P, Casadio R. I-Mutant2. 0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic acids research. 2005; 33(Web Server issue):W306–W310. [PubMed: 15980478]
32. Cheng J, Randall A, Baldi P. Prediction of protein stability changes for single-site mutations using support vector machines. Proteins. 2006; 62:1125–1132. [PubMed: 16372356]
33. Masso M, Vaisman II. Accurate prediction of stability changes in protein mutants by combining machine learning with structure based computational mutagenesis. Bioinformatics. 2008; 24:2002–2009. [PubMed: 18632749]
34. Montanucci L, Fariselli P, Martelli PL, Casadio R. Predicting protein thermostability changes from sequence upon multiple mutations. Bioinformatics. 2008; 24:I190–I195. [PubMed: 18586713]
35. Wu LC, Lee JX, Huang HD, Liu BJ, Horng JT. An expert system to predict protein thermostability using decision tree. Expert Systems Appl. 2009; 36:9007–9014.
36. Gromiha MM, Oobatake M, Sarai A. Important amino acid properties for enhanced thermostability from mesophilic to thermophilic proteins. Biophys Chem. 1999; 82:51–67. [PubMed: 10584295]
37. Huang LT, Gromiha MM. Reliable prediction of protein thermostability change upon double mutation from amino acid sequence. Bioinformatics. 2009; 25:2181–2187. [PubMed: 19535532]
38. Singh RK, Tropsha A, Vaisman II. Delaunay tessellation of proteins: four body nearest-neighbor propensities of amino acid residues. J Comp Biol. 1996; 3:213–221.
39. Zhang Y. Protein structure prediction: when is it useful? Curr Opin Struct Biol. 2009; 19:145–155. [PubMed: 19327982]
40. Wu S, Skolnick J, Zhang Y. Ab initio modeling of small proteins by iterative TASSER simulations. BMC Biol. 2007; 5:17. [PubMed: 17488521]
41. Zhang Y. Template-based modeling and free modeling by I-TASSER in CASP7. Proteins. 2007; 69(Suppl 8):108–117. [PubMed: 17894355]
42. Zhang Y. I-TASSER: fully automated protein structure prediction in CASP8. Proteins. 2009; 77(Suppl 9):100–113. [PubMed: 19768687]
43. Li Y, Middaugh CR, Fang J. A novel scoring function for discriminating hyperthermophilic and mesophilic proteins with application to predicting relative thermostability of protein mutants. BMC Bioinformatics. 2010; 11:62. [PubMed: 20109199]
44. Becktel WJ, Schellman JA. Protein stability curves. Biopolymers. 1987; 26:1859–1877. [PubMed: 3689874]
45. Li YQ, Fang JW. Distance-dependent statistical potentials for discriminating thermophilic and mesophilic proteins. Biochem Biophys Res Commun. 2010; 396:736–741. [PubMed: 20451495]
46. Kryshtafovych A, Krysko O, Daniluk P, Dmytriv Z, Fidelis K. Protein structure prediction center in CASP8. Proteins-Struct Funct Bioinformatics. 2009; 77:5–9.
47. Moult J, Fidelis K, Kryshtafovych A, Rost B, Hubbard T, Tramontano A. Critical assessment of methods of protein structure prediction—round VII. Proteins-Struct Funct Bioinformatics. 2007; 69:3–9.
48. Wu S, Zhang Y. LOMETS: a local meta-threading-server for protein structure prediction. Nucleic Acids Res. 2007; 35:3375–3382. [PubMed: 17478507]
49. Zhang Y, Skolnick J. SPICKER: a clustering approach to identify near-native protein folds. J Comput Chem. 2004; 25:865–871. [PubMed: 15011258]
50. Li Y, Zhang Y. REMO: A new protocol to refine full atomic protein models from C-alpha traces by optimizing hydrogen-bonding networks. Proteins-Structure Function and Bioinformatics. 2009; 76:665–676.
51. Zhang Y. I-TASSER server for protein 3D structure prediction. BMC Bioinformatics. 2008; 9:40. [PubMed: 18215316]

52. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins. 2004; 57:702–710. [PubMed: 15476259]
53. Kumar MD, Bava KA, Gromiha MM, Prabakaran P, Kitajima K, Uedaira H, Sarai A. ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. Nucleic Acids Res. 2006; 34(Database issue):D204–D206. [PubMed: 16381846]
54. Li Y, Drummond DA, Sawayama AM, Snow CD, Bloom JD, Arnold FH. A diverse family of thermostable cytochrome P450s created by recombination of stabilizing fragments. Nat Biotechnol. 2007; 25:1051–1056. [PubMed: 17721510]
55. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990; 215:403–410. [PubMed: 2231712]
56. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983; 22:2577–2637. [PubMed: 6667333]
57. Liang J, Edelsbrunner H, Woodward C. Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. Protein Sci. 1998; 7:1884–1897. [PubMed: 9761470]
58. Masso M, Vaisman II. Accurate prediction of enzyme mutant activity based on a multibody statistical potential. Bioinformatics. 2007; 23:3155–3161. [PubMed: 17977887]
59. Deutsch C, Krishnamoorthy B. Four-body scoring function for mutagenesis. Bioinformatics. 2007; 23:3009–3015. [PubMed: 17921497]
60. Gromiha MM, Huang L-T, Lai L-F. Sequence based prediction of protein mutant stability and discrimination of thermophilic proteins. Lecture Notes Comput Sci. 2008; 5265:1–12.
61. Kim H, Moon EJ, Moon S, Jung HJ, Yang YL, Park YH, Heo M, Cheon M, Chang I, Han DS. New method of evaluating relative thermal stabilities of proteins based on their amino acid sequences; Targetstar. Int J Modern Phys C. 2007; 18:1513–1526.
62. Brych SR, Blaber SI, Logan TM, Blaber M. Structure and stability effects of mutations designed to increase the primary sequence symmetry within the core region of a beta-trefoil. Protein Sci. 2001; 10:2587–2599. [PubMed: 11714927]
63. Culajay JF, Blaber SI, Khurana A, Blaber M. Thermodynamic characterization of mutants of human fibroblast growth factor 1 with an increased physiological half-life. Biochemistry. 2000; 39:7153–7158. [PubMed: 10852713]
64. Dalluge R, Oschmann J, Birkenmeier O, Lucke C, Lilie H, Rudolph R, Lange C. A tetrapeptide fragment-based design method results in highly stable artificial proteins. Proteins-Struct Funct Bioinformatics. 2007; 68:839–849.
65. Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz HR, Ceric G, Forslund K, Eddy SR, Sonnhammer EL, Bateman A. The Pfam protein families database. Nucleic Acids Res. 2008; 36(Database issue):D281–D288. [PubMed: 18039703]
66. Bae E, Bannen RM, Phillips GN Jr. Bioinformatic method for protein thermal stabilization by structural entropy optimization. Proc Natl Acad Sci USA. 2008; 105:9594–9597. [PubMed: 18621726]

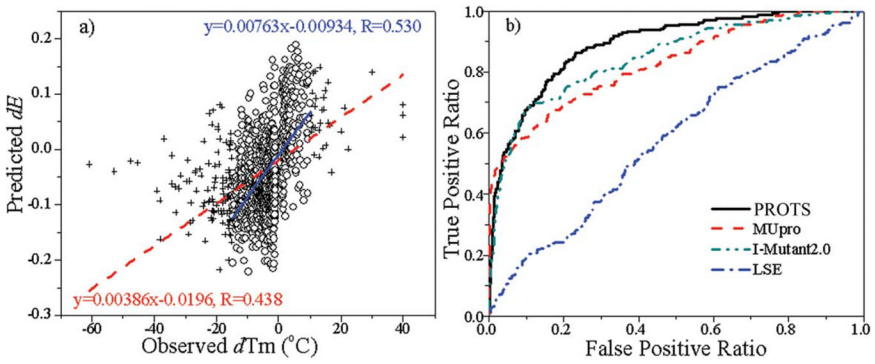

Figure 1.

Linear regression (left) and ROC curves (right) of the 1146 point mutations with ΔT_m values. In the regression plot, the cross points show the mutations with ΔT_m either lower than −15°C or higher than 10°C. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

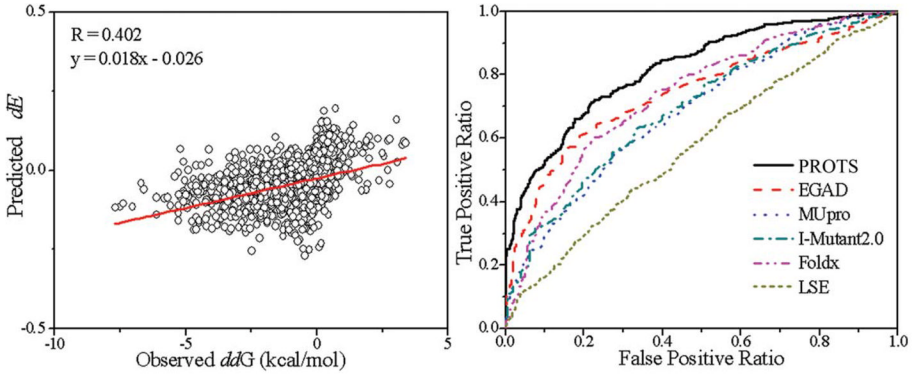

Figure 2.

Linear regression (left) and ROC curves (right) of the 2264 point mutations with ΔΔG measurements. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

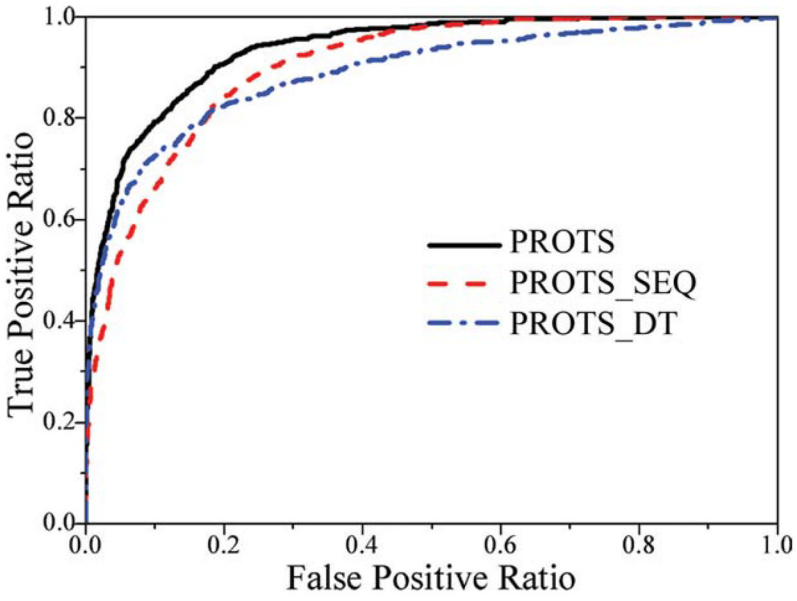

Figure 3.
The ROC curves of PROTS in the classification of 1020 TPs and 4977 MPs. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

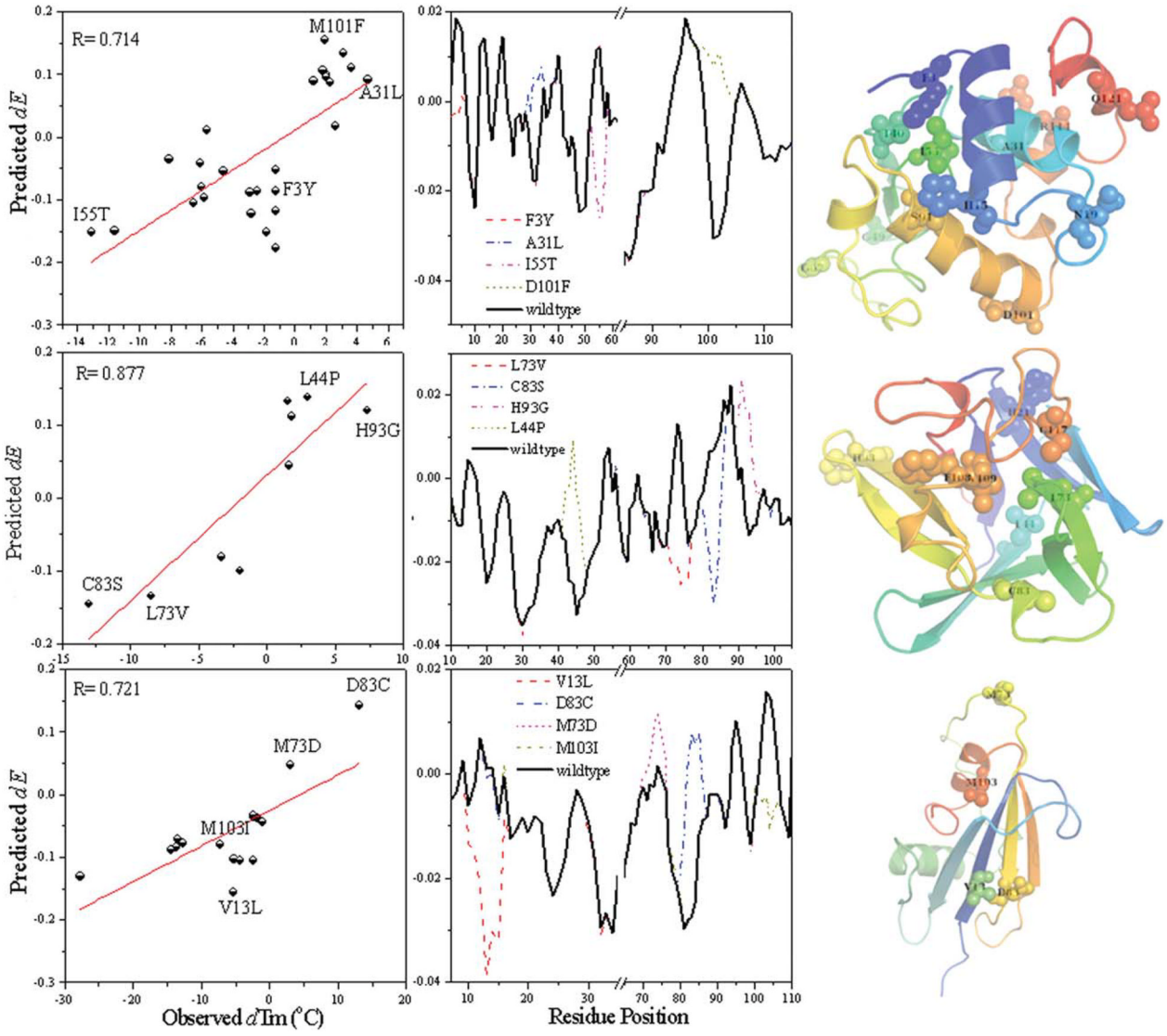

Figure 4.

Examples of PROTS in prediction stability changes for mutants of an alpha-protein (PDBid: 4lyzA, top), a beta-protein (2afgA, middle) and an alpha/beta protein (3ssiA, bottom). The left column presents the regression on all of the mutants. Some significant mutants are labeled. The middle column shows the PROTS potential change at residue level for each mutation. Some unchanged residues are omitted for clarity. The right column illustrates the mutation locations in the wild type proteins. The protein images are generated using Pymol.

Table I

Comparison of ΔT_m Predictions For Mutations and Hypothetical Reversed Mutations

	ALL					Subset[a]				
		WT→MT		MT→WT			WT→MT		MT→WT	
Algorithms	**No. of mutants**	**AUC**	**R**	**AUC**	**R**	**No. of mutants**	**AUC**	**R**	**AUC**	**R**
MUpro[c]	1146	0.828	0.566	0.506	0.063	583[e]	0.643	0.355	0.532	0.099
I-Mutant2.0[b,d]	1146	0.849	0.563	0.558	0.098	502[f]	0.655	0.342	0.545	0.067
LSE	1146	0.578	0.145	0.578	0.145	–	–	–	–	–
PROTS	1146	0.890	0.438	0.890	0.438	1014	0.882	0.530	0.882	0.530
PROTS_SEQ	1146	0.884	0.419	0.884	0.419	1014	0.878	0.514	0.878	0.514

[a] This subset includes the mutations with their ΔT_m values within the range of [−15°C, 10°C]. For MUpro and I-Mutant2.0, the identical mutations included in their training set are also excluded.

[b] The wild type protein 1lrp has only Cα coordinates, so its I-Mutant2.0 predictions are sequence-based.

[c] The results shown here are based on MUpro regression. The AUC from MUpro classification is 0.625 based on ALL mutations and 0.504 based on mutations in the subset.

[d] The results shown here are based on I-Mutant2.0 regression. The AUC from the I-Mutant2.0 classification is 0.686 based on ALL mutations and 0.563 based on mutations in the subset.

[e] The AUC and R of PROTS predictions on the same subset of 583 mutations are 0.878 and 0.408.

[f] The AUC and R of PROTS predictions on the same subset of 502 mutations are 0.869 and 0.444.

Table II

Comparison of the $\Delta\Delta G$ Predictions For Mutations and Hypothetical Reversed Mutations

		AUC		Correlation coefficient (R)	
Methods	**No. of mutants**	**WT→MT**	**MT→WT**	**WT→MT**	**MT→WT**
MUpro	1281[a]	0.687	0.564	0.483	0.167
I-Mutant2.0	933[b]	0.694	0.557	0.540	0.069
LSE	2156[c]	0.577	0.577	0.155	0.155
FoldX[d]	1200[e]	0.738	–	0.497	–
EGAD[d]	1065[f]	0.745	–	0.595	–
PROTS	1500	0.819	0.819	0.402	0.402
PROTS_SEQ	1500	0.815	0.815	0.387	0.387

Mutations identical to the ones used in training were excluded for all algorithms.

[a] For the 1015 forward mutations overlapped in MUpro and PROTS predictions, AUC and R are 0.694 and 0.498 for MUpro, 0.793 and 0.407 for PROTS, respectively.

[b] For the 761 forward mutations overlapped in Imutant2.0 and PROTS predictions, AUC and R are 0.682 and 0.545 for Imutant2.0, 0.773 and 0.306 for PROTS, respectively.

[c] For the 1500 mutations overlapped in LSE and PROTS predictions, LSE presented AUC and R are 0.569 and 0.132.

[d] Prediction values were provided by Dr. Vladimir Potapov.

[e] For the 658 forward mutations overlapped in FoldX and PROTS predictions, AUC and R are 0.692 and 0.448 for FoldX, 0.831 and 0.455 for PROTS, respectively.

[f] For the 779 forward mutations overlapped in EGAD and PROTS predictions, AUC and R are 0.762 and 0.597 for EGAD; 0.823 and 0.438 for PROTS, respectively.

Table III

Comparison of the $\Delta\Delta G$ Predictions For Mutations and Hypothetical Reversed Mutations Using Wild Type and/or Mutant Structures

	140 pairs of single point mutations				All 155 structure pairs			
	R		AUC		R		AUC	
Methods	**WT→MT**	**MT→WT**	**WT→MT**	**MT→WT**	**WT→MT**	**MT→WT**	**WT→MT**	**MT→WT**
MUpro[c]	0.967	0.012	0.971	0.536				
I-mutant2.0[c]	0.940	0.054	0.978	0.534				
PROTS[a]	0.455	0.447	0.840	0.833	0.469	0.463	0.844	0.838
PROTS[b]	0.521	0.521	0.857	0.857	0.574	0.574	0.862	0.862

[a] PROTS values are calculated using wild type protein structures for forward mutations and mutant structures for hypothetical reverse mutations.

[b] PROTS values are calculated using both wild type and mutant protein structures.

[c] MUpro and I-mutants2.0 are not able to predict multiple-mutation induced stability changes.

Table IV

Comparison of PROTS, PROTS_SEQ, FoldX, and LSE in Discriminating 540 TP/MP Orthologous Pairs

Seq. identity	No. of pairs	PROTS	PROTS_SEQ	LSE	FoldX
>30%	345	325 (94.2%)	315 (91.3%)	228 (66.1%)	213 (61.7%)
<=30%	195	190 (97.4%)	183 (93.8%)	139 (71.3%)	102 (52.3%)

The TP/MP pairs are grouped by a threshold of 30% sequence identity to the proteins in the 1020 + 4977 dataset. The number of correct predicted pairs and the accuracies (in parentheses) are shown.

Table V

Comparison of the Applicability of Various Algorithms

Dataset	Algorithms	No. of proteins	No. of proteins with positive correlation	Applicability (%)
The 1146 mutants	MUpro	65	47	72.3
dataset with ΔT_m	I-Mutant2.0	59	41	69.5
values	LSE	78	47	60.3
	PROTS	78	71	91.0
	PROTS_SEQ	78	67	85.9
The 2156 mutants	MUpro	62	42	67.7
dataset with $\Delta\Delta G$	I-Mutant2.0	47	35	74.5
values	LSE	80	49	61.2
	FoldX	59	48	81.4
	EGAD	52	43	82.7
	PROTS	67	55	82.1
	PROTS_SEQ	67	56	83.6

The predictions are grouped by the wild type proteins. The applicability is defined as the ratio of proteins with positive correlation of predicted stability potential changes versus ΔT_m or $\Delta\Delta G$ over all proteins used in the study. An algorithm can be only applicable in proteins with positive correlation.

Table VI

The Value Changes of the PROTS Features in the Mutations 2afgA-H93G and 2afgA-C83S

Features	***dS*(occ)**	***dS*(helix)**	***dS*(strand)**	***dS*(coil)**	***dS*(expose)**	***dS*(bury)**	***dS*(inte)**	***dD*(helix)**	***dD*(strand)**	***dD*(coil)**
2afgA-H93G	0.05081	−0.00289	0.02638	0.01391	−0.03207	−0.00471	0.01364	−0.62592	0.89423	0.73169
2afgA–C83S	−0.07152	−0.01726	−0.00536	−0.02596	−0.02933	−0.01181	−0.01487	0.99053	0.48947	−0.47999
Features	*dD*(expose)	*dD*(bury)	*dD*(inte)	*dS*(occ_DT)	*dS*(D43)	*dS*(D2)	*dS*(D1)	*dD*(D43)	*dD*(D2)	*dD*(D1)
2afgA-H93G	0.84556	−0.06830	0.22273	0.14110	0.20217	0.09523	0.00399	0.00591	0.23480	0.01160
2afgA-C83S	−0.09300	0.71281	0.38019	−0.01118	−0.10821	0.00916	−0.01869	−0.04084	0.30373	−0.24736