Amino Acids (2013) 44:847–855
DOI 10.1007/s00726-012-1407-7

ORIGINAL ARTICLE

Structure-based prediction of the effects of a missense variant on protein stability

Yang Yang · Biao Chen · Ge Tan · Mauno Vihinen · Bairong Shen

Received: 12 June 2012 / Accepted: 23 September 2012 / Published online: 12 October 2012

Abstract Predicting the effects of amino acid substitutions on protein stability provides invaluable information for protein design, the assignment of biological function, and for understanding disease-associated variations. To understand the effects of substitutions, computational models are preferred to time-consuming and expensive experimental methods. Several methods have been proposed for this task including machine learning-based approaches. However, models trained using limited data have performance problems and many model parameters tend to be over-fitted. To decrease the number of model parameters and to improve the generalization potential, we calculated the amino acid contact energy change for point variations using a structure-based coarse-grained model. Based on the structural properties including contact energy (CE) and further physicochemical properties of the amino acids as input features, we developed two support vector machine classifiers. M47 predicted the stability of variant proteins with an accuracy of 87 % and a Matthews correlation coefficient of 0.68 for a large dataset of 1925 variants, whereas M8 performed better when a relatively small dataset of 388 variants was used for 20-fold cross-validation. The performance of the M47 classifier on all six tested contingency table evaluation parameters is better than that of existing machine learning-based models or energy function-based protein stability classifiers.

Keywords Amino acid mutation · Physicochemical properties · Residue–residue contact energy · Support vector machine · Protein stability prediction

Y. Yang and B. Chen contributed equally to this work.

A website with supporting documentation and the software called PPSC (Predictor of Protein Stability Changes) is available at http://www.ibio-cn.org/softwares/PPSC/index.html and http://structure.bmc.lu.se/PPSC/.

Electronic supplementary material The online version of this article (doi:10.1007/s00726-012-1407-7) contains supplementary material, which is available to authorized users.

Y. Yang · B. Chen · G. Tan · B. Shen (✉)
Center for Systems Biology, Soochow University, No1. Shizi Street, Suzhou 215006, Jiangsu, China
e-mail: bairong.shen@suda.edu.cn

Y. Yang
School of Computer Science and Technology, Soochow University, Suzhou 215006, China

G. Tan
Department of Biomedical Engineering, Tongji University, Shanghai 200092, China

M. Vihinen · B. Shen
Institute of Biomedical Technology, University of Tampere and BioMediTech, 33014 Tampere, Finland

M. Vihinen
Department of Experimental Medical Science, Lund University, 221 84 Lund, Sweden

M. Vihinen
Tampere University Hospital, 22521 Tampere, Finland

B. Shen
Bioinformatics Department, Medical College, Soochow University, Suzhou 215123, China

Introduction

Protein stability is a fundamental property affecting function, activity, and regulation of biomolecules. Many amino acid substitutions, which can cause changes to hydrophobicity, over packing, backbone strain, or loss of electrostatic interactions, affect stability and biological functions of proteins, while others are tolerated (Khan and Vihinen 2010). Understanding the effects of these alterations facilitates the elucidation of the molecular basis of many diseases (Thusberg and Vihinen 2009). Site-directed mutagenesis has been utilized for decades (Kearns-Jonker et al. 2007; Rajendhran and Gunasekaran 2007), and several methods have been proposed for predicting protein stability changes based on the primary sequence or the protein's three-dimensional structure information.

Predictions are used because experimental studies are tedious, time-consuming and costly, requiring protein production, purification, and characterization. The predictions methods include energy function-based approaches and machine learning-based methods. The energy functions used in these models include (1) a physical energy function calculated using ab initio quantum mechanics (QM) (Lazaridis and Karplus 2000; Moult 1997), (2) an empirical energy function or force field derived from different types of experimental data (Capriotti et al. 2004), (3) a statistical energy function obtained by the analysis of protein structural information. Both ab initio QM and force field calculations are time-consuming and they are sensitive to small displacements, especially if only a low-resolution protein structure is available. Ab initio QM calculations are impractical for large proteins. Statistical approaches may provide a similar accuracy to the ab initio QM calculation, but their theoretical foundation is not clear (Lazaridis and Karplus 2000).

Recently, several papers have proposed machine learning-based methods, including support vector machines (SVMs) and neural-network (N-N) (Capriotti et al. 2004, 2005a, b; Cheng et al. 2006; Guerois et al. 2002; Shen et al. 2008). Many of these methods are aimed at predicting the sign of $\Delta\Delta G$ (Gibbs free energy change). A positive or negative $\Delta\Delta G$ corresponds to an increase or decrease in the protein stability, respectively. Sequence-based classifiers have been trained and tested using different sequence window lengths (Capriotti et al. 2005a; Cheng et al. 2006), or in combination with structural information (Capriotti et al. 2004). The inputs have been encoded using the 20-alphabet amino acid code, so the number of parameters rapidly rises to more than 100. The large number of parameters may cause problems. The models may be overfitted and give poor performance when applied to new cases (Khan and Vihinen 2010). The database containing experimental $\Delta\Delta G$ values for variations, ProTherm, contains just over 2000 examples (Kumar et al. 2006). The performance of most of the published stability predictors was systematically tested and found to be suboptimal (Khan and Vihinen 2010).

We have previously developed a support vector machine classifier for stability change predictions (Shen et al. 2008). The method uses amino acid similarity from the 20 amino acids based on their physicochemical properties. To further improve the prediction performance, we introduced a new parameter, the protein total contact energy change, which is calculated using the coarse-grained model that we developed (Shen and Vihinen 2003) and we performed systematic optimization of the features for Predictor of Protein Stability Changes (therefore we have termed this method, and software, PPSC). Changes in the total contact energy due to a variation may reflect stability changes. The novel prediction with contact energy change together with the selected physicochemical properties of amino acids utilizes fewer parameters to avoid over-fitting and to improve the performance.

Materials and methods

Datasets

The datasets used to train and test the models was extracted from the ProTherm database (http://gibk26.bio.kyutech.ac.jp/jouhou/Protherm/protherm.html) (Bava et al. 2004; Gromiha et al. 2002). The test cases had to fulfill two criteria: (i) to be single point variations and (ii) the protein structure had to be available in the Protein Data Bank (http://www.rcsb.org/pdb).

We extracted seven attributes for each record including the PDB identification code, the protein name, the amino acid variation, $\Delta\Delta G$, pH and temperature (T) measurements, and the solvent accessibility of the variant residue. If more than one measurement was available for a variant, the average $\Delta\Delta G$ value was used. The final dataset extracted contained 2760 variations affecting 75 proteins, with 1887 negative and 873 positive $\Delta\Delta G$ value cases. A positive $\Delta\Delta G$ value indicates that the variant is stabilizing whereas a negative $\Delta\Delta G$ indicates the destabilizing effect of a variant. This dataset is referred to as S2760.

Any values of $\Delta\Delta G$ between 0.5 and -0.5 kcal/mol were classified as neutral. Therefore, we removed the variations with (absolute value of $\Delta\Delta G$) < 0.5 kcal/mol from S2760 to establish the dataset S1810. The dataset, which contains 1810 variations affecting 71 proteins, with 1,388 negative and 422 positive examples, was used for input attribute analysis due to it having more accurate $\Delta\Delta G$ values.

For comparison to other methods, smaller, previously introduced datasets were used. These included the S1925

set of 1925 single point variants affecting 55 proteins with 1,343 negative and 582 positive cases (Masso and Vaisman 2008), and the S388 set of 388 single site variations in 17 proteins with 340 negative and 48 positive variations (Capriotti et al. 2004).

All the dataset information is shown in Supplementary Table 1. About 200 variants have more than one measurement of $\Delta\Delta G$, and most of the standard deviations of the $\Delta\Delta G$ values for the 200 variants are less than 0.5 kcal/mol (Supplementary Fig. 2). There is less variation within the positive $\Delta\Delta G$ values compared with the negative values. The proportion is broadly 1:2 (except S388) and is caused by the low volume of positive cases contained in the ProTherm database. S388 contains more negative cases, possibly because its variation number is so small. The distribution of the variations in the $\Delta\Delta G$ value in S2760, S1925, and S388 is shown in Supplementary Fig. 1. It accords with normal distribution very well and the data distributed in the section [−4, 4]. The datasets are available at PPSC web sites under the User guide tag.

Support vector machine

SVM is a widely used technique for data classification. The kernel function is used to transform the input space into a higher dimensional feature space for efficient classification (Burges 1998). We built two SVM predictors, M8 and M47, using the radial basis function (RBF) kernel in LIBSVM (www.csie.ntu.edu.tw/∼cjlin/libsvm) (Chang and Lin 2011). M8 has eight and M47 has 47 input vector parameters.

Input attributes for M8

In the construction of the M8 model, eight attributes were used for variant characterization: dHydro, dISA, dElec, dVolume, and dCE, solvent accessibility (ASA) of the variant residue, and pH and T. The attributes dHydro, dISA, dElec, and dVolume were defined as the difference between the original and the variant in amino acid hydropathy (Eisenberg et al. 1984), isotropic surface area (ISA) (Collantes and Dunn 1995), electronic charge concentration (Collantes and Dunn 1995), and volume (Zamyatnin 1972), respectively. The dCE was defined as the difference in total contact energy between the wild type and the variant protein. The first four parameters were calculated as previously described (Shen et al. 2008). The dCE was calculated using our coarse-grained model (Shen and Vihinen 2003). The values of the last three attributes (ASA, pH and T) were extracted from ProTherm.

Of the eight attributes, the first five directly reflect the variant's effect on the structure. The dCE quantitatively describes the total contact energy change. Amino acids are treated as united interacting particles in the coarse-grained model, while the parameters used for the contact energy calculation are environment-dependent and are secondary structure-specific. The coarse-grained model is less sensitive to local structural deviations while it also allows the use of molecular models and relatively low-resolution protein structures (Shen and Vihinen 2003). The solvent accessibility of the amino acid describes the local structural context of a variant and may have different effects depending on whether the variation occurs on the protein surface or deep in the protein core (Ferrer-Costa et al. 2002). By definition, $\Delta\Delta G$ is related to pH and T and thus these two attributes were included.

Input attributes for M47

The M47 model was trained using 47 input attributes, 40 of which are the same as those defined by Capriotti et al. (2005b). The first set of 20 attributes is used to encode for the amino acid types. A value of −1 is given to the wild-type residue, 1 for the variant residue, and 0 for all other amino acid types. The second set of 20 attributes indicates the presence of residue types within a sphere of 9 Å radius. The remaining seven attributes are the same as in the M8 model, i.e., dHydro, dISA, dElec, dCE, ASA of the site, pH, and T.

Assessment of method performance

The goal of the predictor is to forecast the direction of the $\Delta\Delta G$ sign change caused by the amino acid substitutions, without considering the extent of the energy change. The performance of the model was evaluated using the following measures:

$$\text{Accuracy: Acc} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}'}$$

$$\text{True positive rate: TPR} = \frac{\text{TP}}{\text{TP}+\text{FN}'}$$

$$\text{True negative rate: TNR} = \frac{\text{TN}}{\text{TN}+\text{FP}'}$$

$$\text{Positive predictive value: PPV} = \frac{\text{TP}}{\text{TP}+\text{FP}'}$$

$$\text{Negative predictive value: NPV} = \frac{\text{TN}}{\text{TN}+\text{FN}'}$$

Matthews correlation coefficient:

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

where TP, TN, FP, and FN represent the numbers of true positives, true negatives, false positives, and false negatives, respectively.

The performance of the predictors was evaluated with a 20-fold cross-validation, where the dataset was randomly divided into 20 partitions. The model was then trained using 19 of the partitions and tested with the remaining one and repeated 20 times for all combinations before the results were averaged.

We conducted support vector regression (SVR) using the RBF kernel to predict the $\Delta\Delta G$ value. The dependent variable (Y) is the predicted $\Delta\Delta G$ while the experimental $\Delta\Delta G$ is the explanatory variable (X). Correlation coefficient, r, and root mean square error (RMSE) measurements are calculated to assess the prediction performance:

Correlation coefficient:

$$r = \frac{N\sum XY - \sum X \sum Y}{\sqrt{N\sum X^2 - (\sum X)^2}\sqrt{N\sum Y^2 - (\sum Y)^2}}$$

Root mean square error: $RMSE = \sqrt{\frac{\sum_{i=1}(X_i - X_i')^2}{N}}$

The receiver operating characteristic (ROC) curve was calculated to visualize the method performance by plotting the true positive (sensitivity) in comparison with the (1-true negative (specificity)) classification rates. The value of the area under the ROC curve (AUC) is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one when using normalized units (Fawcett 2006; Linden 2006).

Results

Machine learning methods are dependent on the quality and quantity of training data, the training process, and the implementation of the learner. The SVM algorithm can identify with certainty the global minimum of the error (objective) function compared with other methodologies like neural networks (NN); also it requires less computer power when the feature space dimensions increase (Shen et al. 2008).

SVM implementation and parameters optimization

There are four basic kernels of SVM: linear, polynomial, radial basis function (RBF), and sigmoid. Capriotti et al. (2005a) compared the four kernel functions and concluded that the RBF kernel

$$K(x_i, x_j) = e^{-\gamma||x_i - x_j||^2}, \gamma > 0$$

is the first choice in general, because it can deal with non-linear cases, and has at least the same performance as a linear (Keerthi and Lin 2003) or a sigmoid kernel (Lin and Lin 2003) once two parameters, C (penalty parameter which controls the trade-off between the training error and the margin) and γ (the RBF kernel parameter) are optimized. Our previous study (Shen et al. 2008) also suggested that the RBF kernel function performed best in this task; still, we evaluated the four kernel functions with M47 predictor in a 20-fold cross-validation using the dataset S2760. The results (see Fig. 1) indicate that the RBF kernel has the best performance: the accuracy, MCC, and r were 0.85, 0.65, and 0.82, respectively, while the RMSE was the lowest: 1.03 kcal/mol. The polynomial kernel (degree = 3) with the accuracy 0.83 and RMSE 1.18 kcal/mol is the second best relating to performance; the linear and the sigmoid kernels have almost the same performance but not good as that of RBF and polynomial kernels. RBF kernel was used to develop the predictors.

The SVM is optimized by adjusting the parameters C and γ. This was performed using a grid-search method. Figure 2 shows the grid-search result of M8 using the dataset S1810. A Python script from LIBSVM (http://www.csie.ntu.edu.tw/~cjlin/libsvm) (Chang and Lin 2011) called grid.py was automated for the search. The detailed procedure can be found in the Supplementary Materials. A similar grid-search method for support vector regression parameter optimization is supplied by LIBSVM as well with a Python script called gridregression.py. (see Supplementary Table 2 for the optimized parameter values for SVM classification and regression).

Several datasets were used to train and test the methods. Two versions, M8 and M47, were developed. The former was trained with eight physicochemical features for the normal and variant amino acid and the latter with 47 features, the additional ones coding for original and variant amino acid types and types of residues in the vicinity of the site.

Analysis of the input attributes

To test the significance of the input attributes, we produced the S1810 dataset by removing cases with $\Delta\Delta G$ values between 0.5 and -0.5 kcal/mol from the S2760 dataset, because the $\Delta\Delta G$ measurements may have an experimental error of up to $\pm$ 0.4 to 0.5 kcal/mol (Khatun et al. 2004). We trained and tested the predictors using the S1810 dataset. S1810 is the most suitable dataset for this step, as it contains only those cases that clearly affect the $\Delta\Delta G$ value.

Our goal was to determine the parameters that were most important for the protein stability change. We assumed that there was a linear relation between $\Delta\Delta G$ and the eight input attributes and calculated the p value and t-value using R. ASA had the highest statistical significance correlated with the $\Delta\Delta G$ (p value = 1.65e-13 and t-value = 7.432) (Table 1). In addition, dCE, dHydro, dElec, and dISA were all significant (with t-value >2 and p value <0.05) whereas the dVolume and pH had only a marginal correlation.

The effects of the attributes were tested by training the SVM predictors with different combinations of the physicochemical parameters. We found that when we removed any one of the parameters dCE, dHydro, and dISA, the effect on accuracy and MCC was small (Table 2). A bivariate scatter plot of dCE, dHydro, and dISA

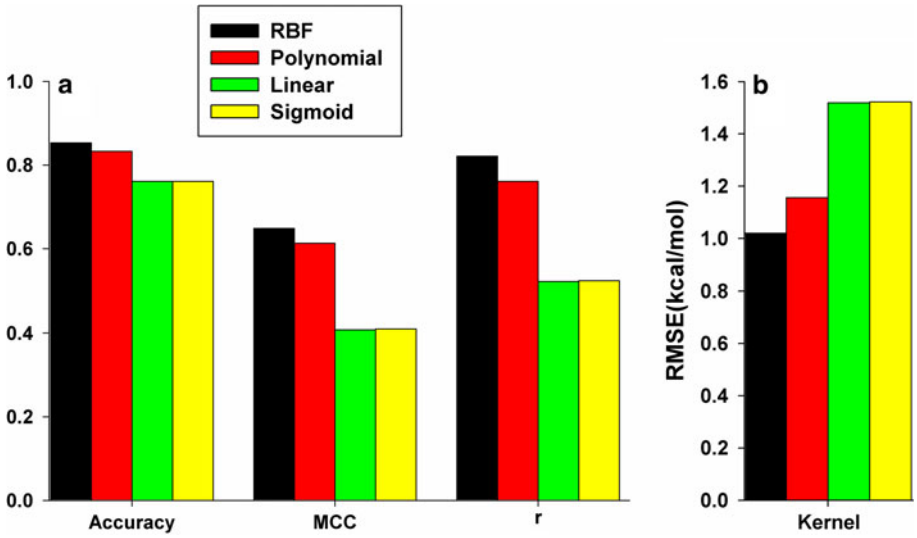

Fig. 1 Performance evaluation of the four kernel functions with our M47 predictor in 20-fold cross-validation using dataset S2760: for each index the 4 *bars* denote RBF, polynomial, linear and sigmoid kernel functions from left to right, **a** the accuracy, MCC, *r*, and **b** the RMSE for the different kernels

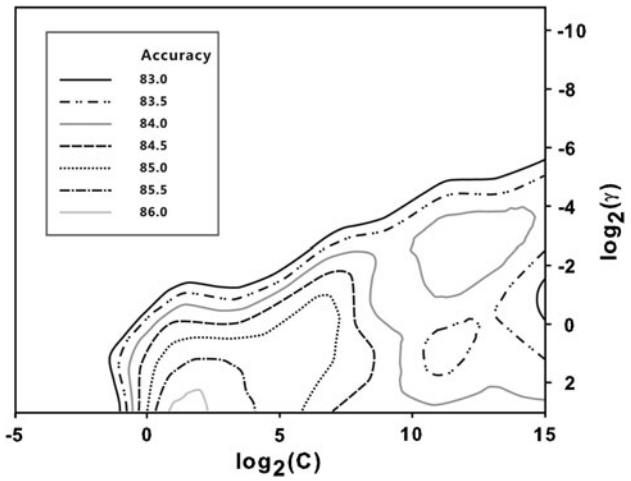

Fig. 2 The contour plot of cross-validation accuracy (in dataset S1810): the accuracies for combinations of the parameters in cross-validation are shown in the legend. The optimal values are $C = 2$, $\gamma = 8$, with CV accuracy of 86.1 %

Table 1 Linear regression analysis of the input attributes

Attributes	Estimate	Std. error	*t*-value	Pr(> |*t*|)*
dCE	0.051	0.015	3.396	0.001
dVolume	−0.004	0.004	−1.094	0.274
dHydro	−0.114	0.257	−4.456	8.87e-06
dISA	−0.007	0.003	−2.469	0.014
dElec	−0.670	0.200	−3.504	4.69e-04
ASA	0.007	0.001	7.432	1.65e-13
pH	0.032	0.023	1.390	0.165
T	0.007	0.002	2.849	0.004

* Pr means *p* values

Table 2 Prediction performance for parameter combinations

Attributes[a]	TPR	TNR	PPV	NPV	Accuracy	MCC
CHEIVAPT	56.8	95.03	77.70	87.87	86.13	0.58
CHEIVA	65.17	91.01	67.57	89.52	85.59	0.56
HEIVA	60.00	91.00	66.58	88.18	83.64	0.53
CEIVA	43.13	96.46	78.79	84.80	84.03	0.50
CHEVA	44.08	96.00	78.48	85.00	84.14	0.51
EIVA	41.46	95.68	74.47	84.72	83.04	0.47
CHEIV	56.16	92.51	69.50	87.41	84.03	0.53
CHEIA	41.7	96.97	80.73	84.55	84.09	0.50

[a] *C* dCE, *H* dHydro, *E* dElec, *I* dISA, *V* dVolume, *A* ASA, *PH* P, *T* Temperature

(Supplementary Fig. 3) shows that these parameters are correlated and when one is removed, the other correlated parameters can compensate for part of the information of the missing parameter. When only limited data are available, increasing the number of parameters may cause the model to be over-fitted. If two of these parameters were removed, the accuracy and MCC declined rapidly. Thus, the dCE, dHydro, and dISA all affect the overall prediction performance. As the combination of all the tested attributes had the best performance it was used as the M8 predictor and with additional features as the M47 predictor, from which dVolume as the least significant physicochemical parameter was excluded.

Prediction performance assessment using the S2760 dataset

M8 and M47 predictors trained with S2760 dataset were tested in 20-fold cross-validation together with FoldX (http://foldx.crg.es/) (Schymkowitz et al. 2005; Guerois

et al. 2002), MUpro (http://www.ics.uci.edu/~baldig/mutation.html) (Cheng et al. 2006) and I-Mutant 2.0 (http://folding.uib.es/i-mutant/i-mutant2.0.html) (Capriotti et al. 2005b). Table 3 shows that M47 outperformed in the comparison. Its accuracy, 0.85 was higher than for any other tool. Overall the performance measures for M47 were greater than those for the other methods and I-Mutant 2.0 comes closest to M47. M8 clearly had a better performance compared with FoldX, but the small number of attributes used for its training is not capable of capturing all the features of the variants.

We conducted regression analysis on all the models and predicted the $\Delta\Delta G$ values for S2760 in the 20-fold cross-validation. Table 4 shows that the r of M47 reached 0.82 while the RMSE was only ± 1.0 kcal/mol, the best for any of the models. Figure 3 shows the relationship between the experimental and the predicted $\Delta\Delta G$ for the M8 and M47 models. The fitted line for M47 is $y = 0.6805x - 0.2494$. The ROC curves in Fig. 4 show the tradeoff between sensitivity and specificity. M47 has the highest performance and I-Mutant 2.0 and Mupro are the closest, whereas M8 and FoldX have clearly weaker performance.

Prediction performance assessment using other datasets

Several other datasets have previously been used to train and test stability predictors. To be able to compare the performance with those methods, further performance evaluations were made. The M8 and M47 methods were further trained with S1925 and S388 datasets.

To study the effects of the training dataset size, we randomly generated ten datasets that contained 10 %, 20–100 % of the dataset S1925. These partitions were then used to train and test in a 20-fold cross-validation the M8 and M47 models as well as two other SVM predictors: I-Mutant 2.0 (with 43 input attributes) and MUpro (with 160 input attributes). When the dataset was small (less than about 300 variants), the prediction accuracy and the MCC of the M8 model were nearly the same as those for M47, and better than for I-Mutant 2.0 and MUpro (Fig. 5, Supplementary Fig. 6). When the size of the dataset was increased, the M47 model was clearly the best. According to this experiment, M8 would be an option to deal with relatively small datasets because M8 has only 8 input vectors, whereas other predictors require more data to learn to generalize.

To compare our results with previously published methods, we trained both M8 and M47 models using the RBF kernel and performed a 20-fold cross-validation with the S1925 dataset. The prediction accuracy of the M47 model was 87 %, the highest among the methods (Supplementary Table 3). The MCC of the M47 model was 0.68, at least 0.05 higher than for the other methods. All the measures were higher for M47 compared with the other methods. The prediction accuracy of M8, which is similar to that for I-Mutant 2.0, is higher than for FoldX, but lower than for MUpro and Automute (http://proteins.gmu.edu/automute/) (Masso and Vaisman 2008).

To determine the $\Delta\Delta G$ value, we trained both M8 and M47 using the RBF kernel and performed a 20-fold cross-validation with the real $\Delta\Delta G$ values in the training dataset. This was implemented using ν-regression SVM with the RBF kernel. With $C = 32$, $\gamma = 0.25$, and $\nu = 0.0625$, the best performance with an r value of 0.84 and RMSE value of 0.97 kcal/mol was obtained using the M47 model (Supplementary Fig. 4, Supplementary Table 4). These results are better than for the other predictors. We also calculated the $\Delta\Delta$G value using FoldX. After optimization and discarding some outliers, FoldX yielded an r value of 0.51 and an RMSE value of 1.8 kcal/mol, the highest among the tested methods. The ROC curves are shown in Supplementary Fig. 5. The AUC values for M8 and M47 are 0.83 and 0.91, respectively, while the AUC for FoldX is 0.78.

M8 and M47 were also assessed using a smaller dataset, S388. When it was used in a 20-fold cross-validation, M8, with its accuracy 0.90, performed even better than M47. However, if S1156, constructed from the S1615 dataset by removing the S388-related dataset, was used for 20-fold cross-validation, the accuracy of M8 was much lower compared with M47 (Supplementary Table 5).

The S388 dataset was also compared with some additional methods, namely MUpro, I-Mutant (http://gpcr2.biocomp.unibo.it/~emidio/I-Mutant/I-Mutant.htm) (Capriotti et al. 2004), DFIRE (http://sparks.informatics.iupui.edu) (Yang and Zhou 2008), PoPMuSiC (version1.0)

Table 3 Prediction performance for the S2760 dataset

Method	TPR	TNR	PPV	NPV	Accuracy	MCC
M8	0.56	0.89	0.70	0.81	0.79	0.48
M47	0.66	0.94	0.84	0.86	0.85	0.65
FoldX	0.64	0.72	0.52	0.81	0.70	0.35
I-Mutant 2.0	0.64	0.93	0.81	0.85	0.84	0.61
MUpro	0.65	0.92	0.79	0.85	0.84	0.61

Table 4 Regression performance for the S2760 dataset

Method	r	RMSE (kcal/mol)	Regression line
M8	0.65	1.3	$y = 0.4676x - 0.4128$
M47	0.82	1.0	$y = 0.6805x - 0.2494$
FoldX	0.41	2.1	$y = 0.4755x - 0.2301$
I-Mutant 2.0	0.79	1.1	$y = 0.5765x - 0.3106$
MUpro	0.78	1.1	$y = 0.6799x - 0.2638$

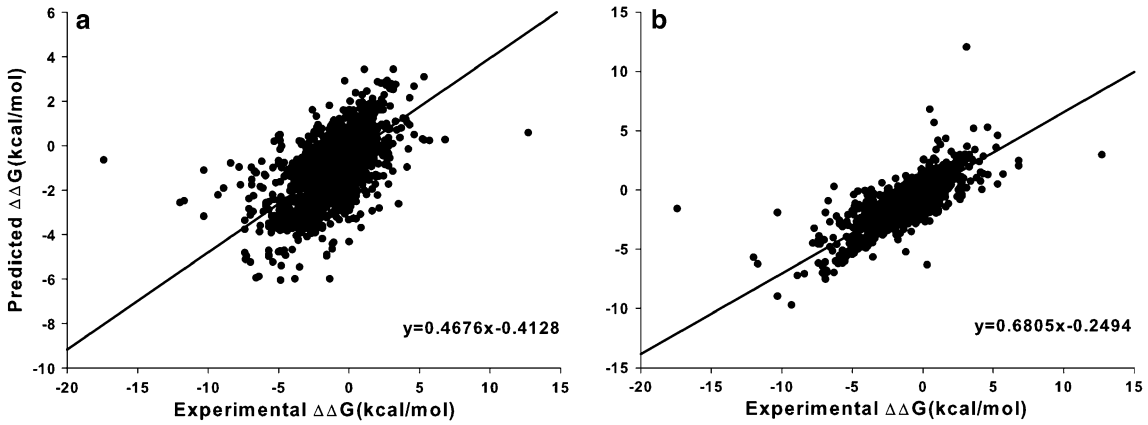

Fig. 3 Regression between the **a** M8 and **b** M47 predicted and experimental $\Delta\Delta G$ using the S2760 dataset

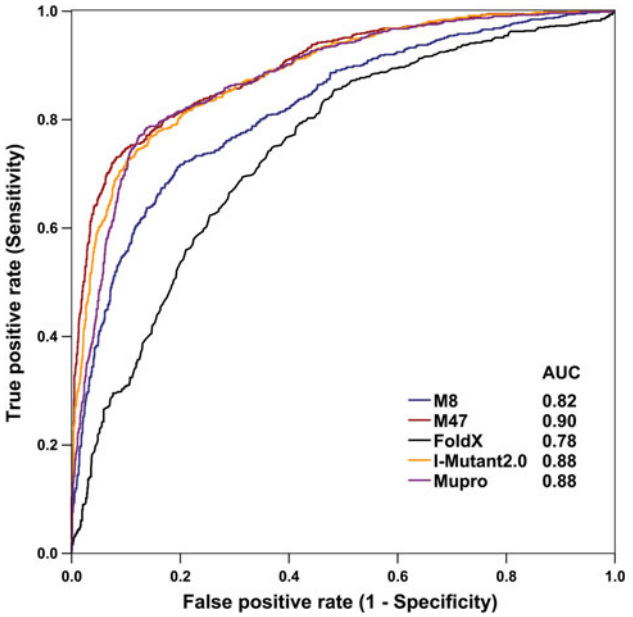

Fig. 4 ROC curves for the M8, M47, FoldX, I-Mutant 2.0, and MUpro models using the S2760 dataset

(http://babylone.ulb.ac.be) (Kwasigroch et al. 2002), and FoldX for which performance figures were available in literature. If S1156 was used as the training dataset and S388 as the blind test set, the accuracy of M47 was 0.89, which was higher than the other datasets (Supplementary Table 6). These results indicate that M47, which has the most descriptive features, provides the best performance.

To compare M47 with our previous SVM predictor (Shen et al. 2008), we used the datasets S2760, S1925, and S1156. The accuracies of the latter were 0.80, 0.79, and 0.86, with MCCs of 0.50, 0.47, and 0.2, respectively. The performance for M47 was significantly improved. Since the previous method does not utilize protein structure information required for all other methods, it is not included in Table 3 or Supplementary Tables 3 or 6.

Prediction performance assessment using different model optimizing methods

Generally, we could apply two different methods to optimize the model parameters: (1) to optimize each of the 20 training sets in the 20-fold CV separately and (2) to optimize the model parameters by optimizing the mean performance over the 20-fold cross-validation. Ideally, if the dataset for model training and parameter optimizing is huge, the two methods will have the same optimizing results, but for many cases of biological studies, with only small datasets available for model building, the models optimized with different datasets will have different results and therefore the model is often unstable.

In this work, we first took the second optimizing method and the results are shown in Supplementary Table 2. For comparison, we also optimized the model using the first method. The optimizing result shown in Supplementary Fig. 7 is similar to the result of second method as expected; although the dataset is not large enough, the results are not exactly the same. With these different models shown in Supplementary Fig. 7, the model cannot be built by averaging the parameters, but only selecting the best model by voting method, in this way, the last model should be the same for these two methods.With $c = 8$, $= 0.5$, the model's performance is indeed the best (see Supplementary Table 3).

Software

The M8 and M47 methods are available in the software tool called PPSC (Prediction of Protein Stability Changes). M8 and M47 models are based on protein structure; thus it is necessary to provide PDB format files when performing prediction. The program can be downloaded from our web page: http://www.ibio-cn.org/softwares/PPSC/index.html or from a mirror site at http://structure.bmc.lu.se/PPSC/.

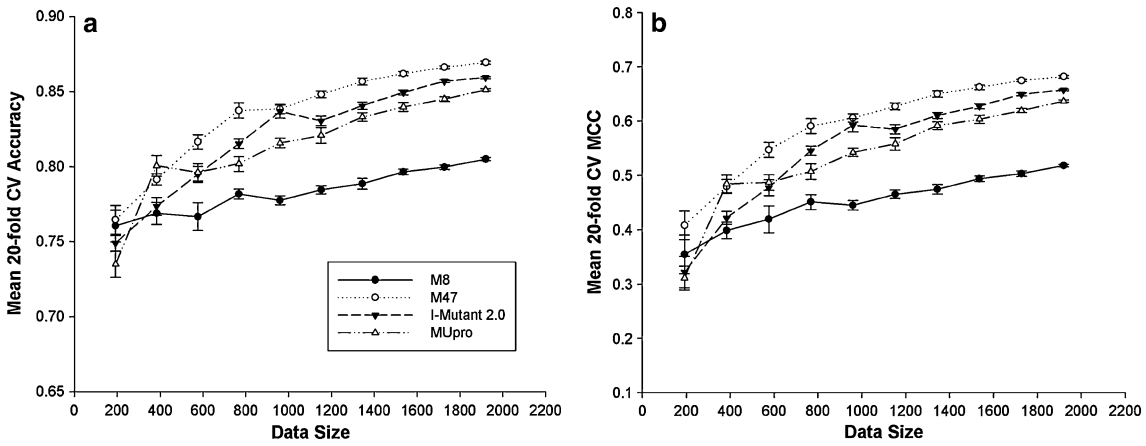

Fig. 5 The classification performance: **a** accuracy and **b** MCC comparison of 4 SVM predictors affected by data size. Tested with different partitions of S1925

When using the program, the first step is to download the PDB file; the CE is then calculated, followed by the stability prediction. There is even an option to train new SVM models with user-provided datasets. PPSC has a provision for batch processing of a list of variants.

Discussion

New methods were developed for protein stability prediction of amino acid substitutions and utilizing approaches, and knowledge from our previous programs (Shen et al. 2008; Shen and Vihinen 2003). The M47 model improves prediction in two main ways compared with previous methods. First, the model quantitatively considers the overall protein structure information using a coarse-grained model, where the contact energy change caused by an amino acid substitution can directly reflect the change in protein stability. The coarse-grained model calculation is rapid because the atomic level information for the amino acids is simplified to united interacting particles. Second, we used only a limited number of parameters for model building (M8). There is a danger of models over-fitting when many parameters and limited datasets are used for model training. However, M47 produced better performance when the dataset size was large. Our predictor imported protein structure information, e.g. the very important information of the structure-neighboring sites, which maybe far away from each other in sequences. Therefore, the improvement benefited prediction accuracy.

The problem of reliably predicting protein stability changes is still important because of its significance when investigating disease mechanisms and during protein engineering. Although several methods have been proposed for predicting stability changes, their performance has not been satisfactory. Our model improved the input features and prediction performance, but it still has two limitations: (1) the dependence on three-dimensional protein structure and (2) the variant modeling is not precise because the native amino acid was replaced with the variant amino acid in the same rotamer. Some software tools can calculate the rotamers for variants including Probe (Word et al. 1999), but due to extensive computational requirements cannot be applied to a large numbers of cases. The coarse-grained model is not sensitive to small deviations in structure and it can compensate for the lack of optimal variant rotamer.

Our future work will focus on studying relationships between input attributes and energy changes to further improve the predictors.

Acknowledgments This work was supported by the National Nature Science Foundation of China (31170795, 20872107), the Specialized Research Fund for the Doctoral Program of Higher Education of China (20113201110015), the Scientific Research Foundation for Returned Scholars, Ministry of Education of China, the International S&T Cooperation Program of Suzhou (SH201120) and the National 973 Programs of China (2010CB945600). The authors gratefully acknowledge the support of K·C Wong education foundation, Hong Kong, the Competitive Research Funding of Tampere University Hospital, Sigrid Juselius Foundation, and Biocenter Finland.

Conflict of interest The authors have declared that no competing interests exist.

References

Bava KA, Gromiha MM, Uedaira H, Kitajima K, Sarai A (2004) ProTherm, version 4.0: thermodynamic database for proteins and mutants. Nucleic Acids Res 32(Database issue):D120–D121. doi:101093/nar/gkh08232/suppl_1/D120

Burges CJC (1998) A tutorial on support vector machines for pattern recognition. Data Min Knowl Disc 2:121–167

Capriotti E, Fariselli P, Casadio R (2004) A neural-network-based method for predicting protein stability changes upon single point

mutations. Bioinformatics 20(Suppl 1):i63–i68. doi:10.1093/bioinformatics/bth92820/suppl_1/i63

Capriotti E, Fariselli P, Calabrese R, Casadio R (2005a) Predicting protein stability changes from sequences using support vector machines. Bioinformatics 21(Suppl 2):ii54–ii58. doi:10.1093/bioinformatics/bti1109

Capriotti E, Fariselli P, Casadio R (2005b) I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res 33(Web Server issue):W306–W310. doi:10.1093/nar/gki375

Chang C-C, Lin C-J (2011) LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol 2(3):1–27. doi:10.1145/1961189.1961199

Cheng J, Randall A, Baldi P (2006) Prediction of protein stability changes for single-site mutations using support vector machines. Proteins 62(4):1125–1132. doi:10.1002/prot.20810

Collantes ER, Dunn WJ 3rd (1995) Amino acid side chain descriptors for quantitative structure-activity relationship studies of peptide analogues. J Med Chem 38(14):2705–2713

Eisenberg D, Schwarz E, Komaromy M, Wall R (1984) Analysis of membrane and surface protein sequences with the hydrophobic moment plot. J Mol Biol 179(1):125–142. doi:0022-2836(84)90309-7

Fawcett T (2006) An introduction to ROC analysis. Pattern Recogn Lett 27:861–874

Ferrer-Costa C, Orozco M, de la Cruz X (2002) Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties. J Mol Biol 315(4):771–786. doi:20015255S0022283601952556

Gromiha MM, Uedaira H, An J, Selvaraj S, Prabakaran P, Sarai A (2002) ProTherm, thermodynamic database for proteins and mutants: developments in version 3.0. Nucleic Acids Res 30(1):301–302

Guerois R, Nielsen JE, Serrano L (2002) Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. J Mol Biol 320(2):369–387. doi:10.1016/S0022-2836(02)00442

Kearns-Jonker M, Barteneva N, Mencel R, Hussain N, Shulkin I, Xu A, Yew M, Cramer DV (2007) Use of molecular modeling and site-directed mutagenesis to define the structural basis for the immune response to carbohydrate xenoantigens. BMC Immunol 8:3. doi:1471-2172-8-3

Keerthi SS, Lin C-J (2003) Asymptotic behaviors of support vector machines with Gaussian kernel. Neural Comput 15(7):1667–1689

Khan S, Vihinen M (2010) Performance of protein stability predictors. Hum Mutat. doi:10.1002/humu.21242

Khatun J, Khare SD, Dokholyan NV (2004) Can contact potentials reliably predict stability of proteins? J Mol Biol 336(5):1223–1238. doi:10.1016/j.jmb.2004.01

Kumar MDS, Bava KA, Gromiha MM, Prabakaran P, Kitajima K, Uedaira H, Sarai A (2006) ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. Nucleic Acids Res 34:D204–D206. doi:10.1093/Nar/Gkj103

Kwasigroch JM, Gilis D, Dehouck Y, Rooman M (2002) PoPMuSiC, rationally designing point mutations in protein structures. Bioinformatics 18(12):1701–1702

Lazaridis T, Karplus M (2000) Effective energy functions for protein structure prediction. Curr Opin Struct Biol 10(2):139–145. doi:S0959-440X(00)00063-4

Lin H-T, Lin C-J (2003) A study on sigmoid kernels for SVM and the training of non-PSD kernels by SMO-type methods. National Taiwan University, Taiwan

Linden A (2006) Measuring diagnostic and predictive accuracy in disease management: an introduction to receiver operating characteristic (ROC) analysis. J Eval Clin Pract 12(2):132–139. doi:10.1111/j.1365-2753.2005.00598.x

Masso M, Vaisman II (2008) Accurate prediction of stability changes in protein mutants by combining machine learning with structure based computational mutagenesis. Bioinformatics 24(18):2002–2009. doi:10.1093/bioinformatics/btn353

Moult J (1997) Comparison of database potentials and molecular mechanics force fields. Curr Opin Struct Biol 7(2):194–199. doi:S0959-440X(97)80025-5

Rajendhran J, Gunasekaran P (2007) Molecular cloning and characterization of thermostable beta-lactam acylase with broad substrate specificity from *Bacillus* badius. J Biosci Bioeng 103(5):457–463. doi:10.1263/Jbb.103.457

Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L (2005) The FoldX web server: an online force field. Nucleic Acids Res 33(Web Server issue):W382–W388. doi:33/suppl_2/W382

Shen B, Vihinen M (2003) RankViaContact: ranking and visualization of amino acid contacts. Bioinformatics 19(16):2161–2162

Shen B, Bai J, Vihinen M (2008) Physicochemical feature-based classification of amino acid mutations. Protein Eng Des Sel 21(1):37–44. doi:10.1093/protein/gzm084

Thusberg J, Vihinen M (2009) Pathogenic or not? And if so, then how? Studying the effects of missense mutations using bioinformatics methods. Hum Mutat 30(5):703–714. doi:10.1002/humu.20938

Word JM, Lovell SC, LaBean TH, Taylor HC, Zalis ME, Presley BK, Richardson JS, Richardson DC (1999) Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms. J Mol Biol 285(4):1711–1733. doi:S0022-2836(98)92400-7

Yang Y, Zhou Y (2008) Specific interactions for ab initio folding of protein terminal regions with secondary structures. Proteins 72(2):793–803. doi:10.1002/prot.21968

Zamyatnin AA (1972) Protein volume in solution. Prog Biophys Mol Biol 24:107–123