Performance of Protein Stability Predictors

Sofia Khan, Mauno A Vihinen

Human Mutation

Performance of Protein Stability Predictors

Journal:	*Human Mutation*
Manuscript ID:	humu-2009-0527.R1
Wiley - Manuscript type:	Informatics
Date Submitted by the Author:	04-Feb-2010
Complete List of Authors:	Khan, Sofia; University of Tampere, Institute of Medical Technology Vihinen, Mauno; University of Tampere, Institute of Medical Technology
Key Words:	protein stability, free energy, missense mutations, stability predictors, prediction programs, bioinformatics, computational methods, predictions

Performance of Protein Stability Predictors

Sofia Khan[1] and Mauno Vihinen[1, 2]*

[1]Institute of Medical Technology, FI-33014 University of Tampere, Finland

[2]Tampere University Hospital, FI-33520 Tampere, Finland

*Corresponding author

Email address: mauno.vihinen@uta.fi

Abstract

Stability is a fundamental property affecting function, activity, and regulation of biomolecules. Stability changes are often found for mutated proteins involved in diseases. Stability predictors computationally predict protein-stability changes caused by mutations. We performed a systematic analysis of eleven online stability predictors' performances. These predictors are CUPSAT, Dmutant, FoldX, I-Mutant2.0, two versions of I-Mutant3.0 (sequence and structure versions), MultiMutate, MUpro, SCide, Scpred, and SRide. As input, 1784 single mutations found in 80 proteins were used, and these mutations did not include those used for training. The programs' performances were also assessed according to where the mutations were found in the proteins, i.e., in secondary structures and on the surface or in the core of a protein, and according to protein structure type. The extents to which the mutations altered the occupied volumes at the residue sites and the charge interactions were also characterized. The predictions of all programs were in line with the experimental data. I-Mutant3.0 (utilizing structural information), CUPSAT, Dmutant, and FoldX were the most reliable predictors, and Scpred was the best of the stability-center predictors. However, at best, the predictions were only moderately accurate (~60%) and significantly better tools would be needed for routine analysis of mutation effects.

Keywords: Protein stability, free energy, missense mutations, stability predictors, prediction programs, bioinformatics, computational methods, predictions

Introduction

Stability is a fundamental property affecting function, activity, and regulation of biomolecules. Conformational changes are required for many proteins' function (Muller et al. 1996; Hsu et al. 2008; Mohamed et al. 2009); therefore, conformational flexibility and rigidity must be finely balanced (Vihinen 1987).

Incorrect folding and decreased stability are the major consequences of pathogenic missense mutations (Bross et al. 1999; Wang and Moult 2001; Ferrer-Costa et al. 2002; Yue et al. 2005). Single residue mutations can cause e.g. reduction in hydrophobic area, over packing, backbone strain, and loss of electrostatic interaction and thus lead to changes in protein stability (Steward et al. 2003). Alterations in atom-atom interactions affect the free energy difference (ΔG) between the folded and unfolded states of proteins. Changes in the interaction among residues within a protein or between a protein and its surroundings affect the entropy of the system with consequent effects in local flexibility/rigidity of the structure (Yue et al. 2005). In addition to covalent disulphide bonds, proteins are stabilized by the noncovalent hydrophobic, electrostatic, and van der Waals interactions, and hydrogen bonds (Pace 1990; Ponnuswamy and Gromiha 1994). Cooperative, noncovalent, long-range interactions provide stability that counteracts local tendencies to unfold (Abkevich et al. 1995; Gromiha and Selvaraj 2004). The importance of the interactions for stability has been revealed by site-directed mutagenesis experiments (Villegas et al. 1996; Akasako et al. 1997; Petsko 2001; Sawano et al. 2008). Intramolecular interactions define the overall structure and stability of a protein, as well as regions that can undergo conformational

rearrangements. Additionally, functions, such as catalysis, allosteric regulation, and ligand binding, depend mostly on the same interactions that define stability.

Understanding the mechanisms by which mutations affect protein stability is an important subject. Accurate prediction of protein stability changes that arise upon mutagenesis is necessary if the structure-function relationship of a protein is to be understood or if a new protein is to be designed. Understanding the structure-function relationship is also essential when characterizing disease mechanisms (Sunyaev et al. 2001; Thusberg and Vihinen 2009) and evolutionary dynamics (Bloom et al. 2005; DePristo et al. 2005; Pal et al. 2006; Bloom et al. 2007; Camps et al. 2007; Poelwijk et al. 2007), and when designing or engineering proteins (Baltzer and Nilsson 2001; Lehmann and Wyss 2001; Bolon et al. 2002; van den Burg and Eijsink 2002; Bloom et al. 2005; Butterfoss and Kuhlman 2006).

Many computational methods have been developed to predict the difference in the free energy of unfolding ($\Delta\Delta G$) between a wild-type protein and its mutant. Some of these methods rely on energy functions to compute the $\Delta\Delta G$, while others apply machine-learning approaches. The methods that use energy functions can be subdivided to: physical potential approaches, statistical potential approaches, and empirical potential approaches (Capriotti et al. 2004). The physical potential approaches (Bash et al. 1987; Prevost et al. 1991; Pitera and Kollman 2000) simulate the atomic force-fields of a structure and cannot therefore be applied to large datasets because they are computationally intense. Statistical potential approaches (Gilis and Rooman 1997; Gilis and Rooman 2000; Zhou and Zhou 2002; Zhou and Zhou 2004; Magyar et al. 2005; Deutsch and Krishnamoorthy 2007) use potential functions derived from statistical

analyses of environmental propensities, substitution frequencies, and correlations of adjacent residues found experimentally in protein structures. For the empirical-potential approach (Guerois et al. 2002; Cheng et al. 2006; Parthiban et al. 2006), the energy function is a combination of the weighted physical and statistical energy terms and structural descriptors. Machine-learning methods (Dosztanyi et al. 1997; Dosztanyi et al. 2003; Capriotti et al. 2005; Cheng et al. 2006; Capriotti et al. 2008; Shen et al. 2008) are first trained using examples of proteins and their mutants for which the ΔΔGs have been experimentally measured. Recently a combination of these approaches has been developed (Masso and Vaisman 2008).

Experimental studies on the molecular effects of mutations are often laborious, time-consuming, and costly. Computational and statistical methods may be used instead to predict many of the effects caused by mutations and to elucidate the underlying biological mechanisms (Thusberg and Vihinen 2009). We performed a systematic analysis of the performances of eleven stability predictors available on the Internet. The developers of these methods have used different datasets to test the accuracies of their programs; therefore; a comprehensive, comparative assessment of their performances has yet to be made. Our analysis revealed that the predictive performances of the methods clearly differ and there is a need for more reliable tools.

Methods

The novel methods that produce vast biological datasets demand bioinformatics tools and methods to analyze and interpret the observations. For certain tasks several tools may be available, but without reliable knowledge about the performance and quality of

predictions choosing the correct tool to use is not possible. We therefore performed a comprehensive evaluation of eleven bioinformatics tools designed to predict protein stability changes.

Test Cases

We built a dataset containing missense mutations for which the corresponding proteins had experimentally determined ΔΔG values from ProTherm database (ProTherm update Dec. 19, 2008) (Kumar et al. 2006). ProTherm is the most comprehensive database for experimentally determined protein stability free energy changes caused by mutations. Mutations with associated ΔΔG values between 0.5 and −0.5 kcal/mol were classified as neutral cases, not affecting stability, because the experimental error for measurement of ΔΔG has been estimated as ±0.48 kcal/mol (Khatun et al. 2004). We defined positive cases as having ΔΔG values ≥0.5 or ≤–0.5 kcal/mol. We did not consider proteins containing double mutations and used only one representative case when several ΔΔG values from different studies were available for a given mutation. The final dataset contained 1784 mutations from 80 proteins, with 1154 positive cases of which 931 were destabilizing (ΔΔG ≥ 0.5 kcal/mol), 222 were stabilizing (ΔΔG ≤ –0.5 kcal/mol), and 631 were neutral (0.5 kcal/mol ≥ ΔΔG ≥ –0.5 kcal/mol). (Note that the signs for the ΔΔG values are the opposite those given in the ProTherm database.)

The sizes of the datasets used to test the stability predictors varied, because the majority of the predictors had been trained using data obtained from earlier versions of ProTherm; therefore, only those cases that had been added to the database after training had occurred were used. The datasets for I-Mutant2.0, CUPSAT, FoldX, Dmutant, and

MultiMutate included 174, 536, 1541, 1714, and 1757 mutations, respectively. The smallest datasets used that contained enough cases for statistical analysis was for MUpro (166 mutations) and both versions of I-Mutant3.0 (115 cases each). For the programs SCide, SRide, and Scpred, which predict the existence of stability centers, the datasets contained 1646, 1589, and 1784 mutations, respectively. For AUTO-MUTE, the dataset contained only 28 cases.

Prediction Methods

The effects of mutations on protein stabilities were predicted using the default parameters of the programs were always used. We ran the programs at the Pathogenic-or-Not Pipeline (Thusberg and Vihinen 2009). This service submits the input data, i.e., the wild-type protein structure and/or sequence, and the amino acid substitution, to the selected predictors and parses the results of the individual methods into a single output.

AUTO-MUTE (Masso and Vaisman 2008) (http://proteins.gmu.edu/automute/AUTO-MUTE.html) uses a four body, knowledge-based, statistical contact-potential. The program calculates an empirical, normalized measure of the environmental perturbation for substitutions. A feature vector is used to estimate the effect of the mutation by considering the spatial perturbation inflicted by the mutation upon its nearest neighbors in the 3D structure. We used the random forest option.

CUPSAT (Parthiban et al. 2006) (http://cupsat.uni-koeln.de) predicts $\Delta\Delta G$ using structural, environment-specific, atomic potentials and torsion-angle potentials derived from non-redundant protein structures (Wang and Dunbrack 2003). The torsion-angle potentials are derived from the distribution of protein backbone φ and ψ angles in the dataset.

Dmutant (Zhou and Zhou 2002) (http://sparks.informatics.iupui.edu/hzhou/mutation.html) uses a statistical potential approach with a distance-dependent, residue-specific, all-atom, and knowledge-based potential for protein structure-based predictions.

FoldX version 3.0 (Guerois et al. 2002) (http://foldx.crg.es/) is an empirical potential approach that uses an energy function derived from a weighted combination of physical-energy terms, statistical-energy terms, and structural descriptors calibrated to fit experimental ΔΔG values. FoldX and Dmutant are the only programs discussed herein that return negative ΔΔG values for stabilizing mutations and positive values for destabilizing mutants.

I-Mutant2.0 (Capriotti et al. 2005) (http://gpcr2.biocomp.unibo.it/cgi/predictors/I-Mutant2.0/I-Mutant2.0.cgi) and I-Mutant3.0 (Capriotti et al. 2008) (http://gpcr2.biocomp.unibo.it/cgi/predictors/I-Mutant3.0/I-Mutant3.0.cgi) are support vector machine (SVM)-based tools. The services use either a protein structure or a sequence as input. We used the sequence-based version of both the versions as well as the structure based version of I-Mutant3.0. I-Mutant2.0 programs can be used to predict the sign of the stability change upon mutation or as a regression estimator to predict ΔΔG values. Unlike other stability predictors analysed here, the I-Mutant3.0 classifies the prediction in three classes: neutral mutation (-0.5=<ΔΔG=<0.5), large Decrease (< -0.5) and large Increase (> 0.5).

MultiMutate (Deutsch and Krishnamoorthy 2007) (http://www.math.wsu.edu/math/faculty/bkrishna/DT/Mutate/) uses a four-body scoring function based on Delaunay tessellation of proteins. The method calculates the change

in how well packed the residues are in the wild-type protein and in the mutant. Score values between 0.5% and –0.5% are classified as negative.

MUpro version 2.0.4 (Cheng et al. 2006) (http://www.igb.uci.edu/servers/servers.html) contains two machine-learning programs, SVM and Neural Networks. We used the sequence-based version of the program. The SVM method was run using the default parameters. The output of the program is the sign of the energy change (+ or –).

The programs SCide (Dosztanyi et al. 2003), Scpred (Dosztanyi et al. 1997), and SRide (Magyar et al. 2005) identify stability centers from sequence data. Mutations found at stability centers were considered by us to be destabilizing and thus deleterious. SCide (http://www.enzim.hu/scide) attempts to identify stability centers within experimentally determined protein structures. Stabilizing, cooperative, long-range contacts identified by SCide are formed between regions that are sequentially well separated or that are part of different subunits within a complex. Scpred (http://www.enzim.hu/scpred/pred.html) locates stability-center elements that impart stability via cooperative, long-range interactions. Scpred uses a neural network to predict stabilizing residues in conjunction with sequence information for the protein under study and its homologues. SRide (http://sride.enzim.hu/) combines several methods to identify residues expected to play key roles in stabilization. It analyzes tertiary structures, rather than primary structures, and the evolutionary conserved residues contained within. A residue is predicted to be stabilizing if it is surrounded by hydrophobic residues, exhibits long-range order, has a high conservation score, and, if it is part of a stability center.

Determination of Protein Structural Classes for the Test Cases

CATH (class, architecture, topology, homology; http://www.cathdb.info/), a hierarchical protein-domain classification system (Orengo et al. 1997), was used to group the proteins according to secondary structure type and tertiary organization (protein structure type).

Determination of Secondary Structural Elements and Accessible Surface Areas

Secondary structural information for each mutation site was obtained from ProTherm where the data is taken from PDB file annotations. Accessible surface area (ASA) values were obtained from ProTherm, originally computed using the program, Analytical Surface Calculation. We classified residues with <10% ASAs as buried and with >25% ASAs as exposed.

Determination of Volume and Charge Changes

To calculate the residue-site charge and volume changes that would occur upon mutation, we obtained from the literature amino acid isoelectric point values (Greenstein and Winitz 1961) and volumes (Pontius et al. 1996).

Statistical Analyses

In the analysis the net effect i.e. the sign of the predictions was used. The ΔΔG values were used only to separate neutral cases from positive ones. The quality of the predictions is described by four parameters. In the following equations, *tp*, *fp*, *tn*, and *fn* refer to the number of true positives, false positives, true negatives, and false negatives, respectively.

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}$$

$$\text{Specificity} = \frac{tp}{tp + fp}$$

$$\text{Sensitivity} = \frac{tp}{tp + fn}$$

$$\text{MCC} = \frac{tp \times tn - fn \times fp}{\sqrt{(tp + fn)(tp + fp)(tn + fn)(tn + fp)}}$$

Matthew's correlation coefficients (MCC) range from −1 to 1. A value of MCC = 1 defines the best possible prediction, while MCC = −1 indicates the worst possible prediction (or anti-correlation). For MCC = 0, the prediction is the result of chance. To be able to correlate the quality parameters for different programs with different sizes of test sets containing different amounts of positive and negative cases, the numbers of negative cases were normalized to be equal to the number of positive cases for each program. We used receiver operating characteristics (ROC) curves to plot the balance between sensitivity and specificity. ROC analysis was run at http://www.jrocfit.org.

Mutation statistics were analyzed by comparing the frequencies of the mutations with the expected values that were calculated using the distribution of all amino acids in the analyzed dataset. For the mutated residues, the expected values were calculated with regard to their codon diversity thereby taking into account all possible amino acid substitutions.

The χ^2 test was used to determine the significance of the results and χ^2 was calculated as:

$$\chi^2 = \Sigma \frac{(f_o - f_e)^2}{f_e}$$

where f_o is the observed frequency and f_e is the expected frequency for an amino acid. *P*-values were estimated in a one-tailed fashion.

Correlations between the program outputs were calculated by counting all of the common cases and those predicted correctly.

Results

The performances of the eleven stability predictors differed when tested with our ProTherm dataset. SCide (Dosztanyi et al. 2003) and Scpred (Dosztanyi et al. 1997), which predict stability centers, as well as SRide (Gromiha and Selvaraj 2004), which predicts stabilizing residues, can predict only destabilizing effects caused by mutations. The other programs evaluate both stabilizing and destabilizing changes.

Fig. 1A diagrams the distributions of the predicted and the experimental ΔΔG values follow normal distribution curves. The values predicted by I-Mutant2.0 and CUPSAT are somewhat biased towards negative values, whereas, those for Dmutant trend towards positive values, although the highest peak in the curve for the Dmutant data is at ΔΔG = 0. The distribution for the FoldX results does not show a clear peak; however, there is a peak at the negative end, and many of the ΔΔG values predicted by FoldX are smaller than -4 kcal/mol.

To evaluate the performances of the programs, we used four measures: accuracy, specificity, sensitivity, and MCC. Table 1 displays the values of these measures for all of the mutations and individually for the stability-increasing and -decreasing mutations. The overall performances are best for I-Mutant3.0 (structure version), Dmutant and FoldX, which all have accuracies varying from 0.54 to 0.64. MUpro returned the best

sensitivity value (0.74); while for I-Mutant2.0 and CUPSAT, the values are only slightly smaller (0.71 and 0.69, respectively). The specificity (0.63) is best for I-Mutant3.0 (structure version). However, the MCC values are poor for all the predictors, the best being I-Mutant3.0 (structure version) that has MCC of 0.27. The worst overall MCC value (-0.39) was obtained for MUpro.

All the programs succeed better when considering their ability to predict stability increasing or decreasing mutations individually. In these analyses only two classes were considered, stabilizing or destabilizing and neutral cases. The neutral cases thus contained also destabilizing or stabilizing cases, as well, depending on the analysis. CUPSAT has the highest accuracy, sensitivity and MCC for stabilizing mutation predictions, 0.74, 0.43 and 0.35, respectively. Due to low number of stabilizing cases (5) among I-Mutant3.0 datasets, they were excluded. I-Mutant3.0, FoldX and Dmutant are the best methods for the prediction of destabilizing mutations all having MCC around 0.38. Sensitivity measures the proportion of true positive cases that are correctly identified. MUpro and I-Mutant3.0 (sequence version) has the best sensitivity values. All the programs have specificity over 0.50. Of the stability-center predictors, which only predict destabilizing mutations were equally accurate, but on other terms Scpred was the most reliable and SRide was the poorest predictor. The results for these programs are somewhat poorer than for the best general predictors. The ROC curves for the performances of FoldX, I-Mutant2.0, Dmutant, and CUPSAT are shown in Figure 1B. The steep increase in the curves indicates that these programs were all capable of predicting the stability effects caused by the mutations. However, the curves bend strongly already at tp ~0.6. The AUCs for these programs are between 0.79 and 0.83.

Analysis of Structural Properties

The effects that the type of mutation had on prediction performance were tested by determining the number of times a mutation replaced or substituted for a given amino acid, occurred within a secondary structural element or within a protein folding type, and caused a change in residue size or charge. The distributions of the original (mutated) and substituted (mutant) residues are given in Supplementary table 1. Among the mutated residues that are replaced by stabilizing mutations, D and H are significantly overrepresented, and P and K are significantly underrepresented. Among the mutated residues that were replaced by ones causing destabilization, C, I, and V are significantly overrepresented, while E, G, K, Q and S are significantly underrepresented. For residues replaced by mutations that changed |ΔΔG| by 0.5 kcal/mol or less (neutral mutations), the distributions are also biased but involve different residues. Mutations to P, G and L are much rarer than expected, while E, D, and V are overrepresented. Among the mutant residues, the distributions are even more biased. For all categories, but particularly those involving destabilizing or neutral mutations, alanine substitutions are greatly overrepresented. This observation contradicts the basic assumption behind alanine-scanning mutagenesis (Cunningham and Wells 1989), i.e., alanine substitutions are assumed to affect only the function of the substituted residue (and not the stability of the protein). Destabilizing alanine substitutions were found mainly in coils, turns, and β-strands (33× greater than expected for coils, 26.3× greater for β-strands, and 15.5× greater for turns, when compared with the wild-type alanine distribution). The mutation profiles are clearly

different for stabilizing and destabilizing mutations. The distribution for stabilizing mutant residues is nearly random.

The results for the mutations in the secondary structural elements are given in Fig. 2A. The dataset for I-Mutant3.0 was too small. Overall, the majority of the programs predict different secondary structural elements with almost equal accuracy. CUPSAT predicted, with somewhat better accuracy than did the other programs, the effects of mutations that occurred in coils and turns. For all structural categories, I-Mutant2.0, FoldX, MUpro, MultiMutate, and CUPSAT gave the best results for sensitivity. When accuracy, specificity and sensitivity were considered, Dmutant performed better for mutations found in α-helices and coils and performed poorly for mutations in strands or turns. FoldX, I-Mutant2.0 and MultiMutate are predicting different secondary structures with almost equal specifity, whereas other predictors have differences in this respect. Proteins are classified by CATH as mainly α-helical, as mainly β-stranded, as mixed α and β structures, or as having few secondary structures. The predictions obtained from the eleven programs differed with respect to performance depending on which protein class type a mutation was found in (Fig. 2B). CUPSAT, Dmutant, FoldX, I-Mutant2.0 and MultiMutate made the most accurate and sensitive predictions for mutations that are in domains or proteins composed of few secondary structures. All programs showed great variability in specificity when different protein structure types were compared, e.g., I-Mutant2.0 predicted the effects of mutations in β-strand proteins with an accuracy of 0.34, in α and β proteins with an accuracy of 0.53, and in α-helical proteins with an accuracy of 0.84. The predictive specificities of MultiMutate and Scpred vary only slightly for the different protein structure types. Additionally, the MCCs for the

programs deviate widely. Five out of eight programs (MUpro lacks the respective value) have highest MCC for proteins composed of few secondary structures.

Often, a mutation, associated with a disease state, drastically changes the chemical and/or physical properties at the mutated site. One such change is a change in the accessible surface area (ASA). We considered residues with ASA values of at least 25% those of fully exposed amino acids to be surface residues and those having ASA values of ≤ 10% to be buried. All programs, except MultiMutate, predict exposed mutations more accurately than buried mutations (Fig. 2C). There are major categorical differences in prediction sensitivity for CUPSAT, Dmutant, FoldX, Multimutate and MUpro. Predictions for mutations among buried residues are more specific than for amino acids on surface except for MultiMutate. All programs predicted the effects that the buried mutations had on stability with more accuracy and specificity than they did the stability effects associated with surface residues.

The performances of the predictors as a function of volume change upon mutation are shown in Supplementary Fig. 1. When the original residue is replaced with a residue of smaller volume, a cavity may form in the protein interior. Large volume changes were predicted better than were small changes by all the programs. In comparison with the experimental data, the distributions of correct predictions are similar for CUPSAT and MultiMutate. The distributions of the false positives for the stabilizing mutations are all quite similar except that the peak positions do not coincide. The distributions of destabilizing mutations predicted by the programs follow the experimental distribution very closely. For the false positive distributions, that produced by Scpred differs substantially from the others. The performances of the predictors were unbiased with regard to the type of mutation and the accuracy of the prediction.

The distributions caused by changes in charge are presented in Supplementary Fig. 2. For destabilizing mutations there are no significant performance deviations in the methods for different charge changes. The results obtained using I-Mutant2.0 and MUpro are not reliable because only eight mutations within their datasets changed charge. The distributions obtained for the neutral cases are similar to those found for the experimental data, except for those of the Scpred and MultiMutate. In summary the predictors performed similarly despite differences in the extent to which the volume or charge varied as functions of the original residue and the mutation.

To further assess the performances of the programs we compared the predictions obtained for the same mutations used by the programs in a pairwise fashion (Table 2). The programs were tested with different datasets, which avoided using the training cases. The most similar test sets were for Scpred and MultiMutate, which shared 98.5% of the cases. Conversely, the dataset used for the CUPSAT and I-Mutant2.0 comparison had only 18 mutations (1% of the original dataset). The largest percentage of correctly predicted cases was 38% (for the Dmutant and I-Mutant2.0 comparison). On average, the number of correctly predicted cases was less than one-third of the total data in each set. The correlation between two programs was best for MUpro and SRide, relatively good for SCide and SRide and for CUPSAT and MUpro, and the worst for SRide and I-Mutant2.0. In general however, the overall performances varied greatly because the correlations between programs were found to be small.

Figure 6 shows the agreement among the programs with the experimental data. For the vast majority of cases when only the six general methods were considered, the predictions of just one to three of the methods are in agreement, and when all eleven predictors were considered, only one to four of the predictions agree. There was not a

single case for which all of the programs correctly predicted the experimental result, and when only the general predictors were considered together, in 16% none of their results agree with the experimental data.

Discussion

We evaluated how reliably the stability effects of missense mutations could be predicted. Stability changes can be studied experimentally, but such studies are laborious, time consuming, and often costly. Therefore, reliable computational methods that can predict stability changes are valuable tools. Mutations that decrease the stability of proteins are generally considered to be harmful. In some circumstances, mutations that increase protein stability can also be deleterious. Proteins are dynamic molecules, and mechanical flexibility is necessary for their function (Vihinen 1987; Fields 2001; Daniel et al. 2003). Increased stability can reduce flexibility (Somero 1995; Wolf-Watz et al. 2004). The active-site residues of enzymes are generally polar or charged, and are usually located in hydrophobic clefts (Fersht 1999). Stabilizing mutations in active site residues can reduce enzymatic activities (Zhi et al. 1991; Meiering et al. 1992; Schreiber et al. 1994; Kidokoro et al. 1995; Shoichet et al. 1995; Garcia et al. 2000; Beadle and Shoichet 2002; Mukaiyama et al. 2006; Nagatani et al. 2007; Counago et al. 2008). Additionally, a stabilizing mutation increased the resistance of ribonuclease A to proteolysis, (Markert et al. 2001), which, for example, would be an undesirable effect if it occurred in enzymes involved in cell signaling (Fink 2005).

We tested the performances of eleven protein stability predictors. For this study, we used only sequence data as input for I-Mutant2.0, MUpro, and Scpred, even though the first two programs can also use structural information. CUPSAT, Dmutant, MultiMutate, SCide, and SRide require structural information as input data. Bioinformatic studies concerning protein stability predictions have often used tertiary structure information, because such information has improved the quality of the predictions and, indeed, we found that CUPSAT, Dmutant, and FoldX were the best of the predictors. However, even though Scpred uses only sequence data as input, it returned the most accurate predictions among the stability-center predictors. Although there are two versions of MUpro—one that uses structural and sequence data and one that uses only sequence data—the two versions of the program are quite similar (Cheng et al. 2006) and therefore, we used the sequence-based version.

Certain aspects of the performance of stability predictors have been tested in three previous studies. Potapov and colleagues (Potapov et al. 2009) compared the performances of six programs, CC/PBSA, EGAD, FoldX, I-Mutant2.0, Rosetta, and Hunter. I-Mutant2.0 and FoldX are the only predictors also used in our study. Their dataset was composed of 2156 single mutations obtained from ProTherm. As with our study, mutations that were used to train the programs were not used in their trials. None of the programs they assessed performed as well as reported by their developers, which is what we also found. Of the tested programs, EGAD (Pokala and Handel 2005) cannot predict effects for all types of mutations, and a description of Hunter has not been published and the program is not available. We identified web services that could be used in conjunction with only sequence data, mutation positions, and, in some cases, coordinates of the wild-type protein as input, and then used those services without

subsequent user intervention. CC/PBSA (Benedix et al. 2009) did not meet these criteria, as it requires the use of two programs and extensive computing power. Rosetta software is used for protein modeling and design. The intent of Potapov et al. was to correlate experimental and predicted ΔΔG values, while we were interested in determining whether the stabilizing or destabilizing effect caused by a mutation could be correctly predicted, because, for mutations associated with disease states, the sign of the stability change is what is needed.

Lonquety and colleagues (Lonquety et al. 2008) evaluated predictors that detect folding nuclei affected by mutations. The programs tested included Dmutant, the two versions of I-Mutant2.0, MUpro, and PoPMuSiC. Their dataset contained 1409 mutations from the ProTherm. However, they tested I-Mutant2.0 and MUpro with same dataset that had been used for training. Thus, their results indicated only how well the methods learned the training set. The correlation coefficients for PoPMuSiC and Dmutant were ~0.5. We did not test PoPMuSiC because the server for the version available at the time was very unstable. A new, more stable version (Dehouck et al. 2009) was released after we finished our study. We could not test the newer version because its neural network was trained using a more current set of ProTherm data, and thus, there were not enough test cases available.

Tastan and colleagues (Tastan et al. 2007) used three structure-based programs, Dmutant, FoldX, and I-Mutant2.0, to investigate stability predictions for mutations in two types of membrane proteins, mammalian rhodopsins (279 mutations) and bacteriorhodopsins (54 mutations). The best prediction accuracy for the rhodopsin dataset was <0.60, while it was somewhat greater for the bacteriorhodopsin dataset.

Only 20% of the rhodopsin dataset and 35% of the bacteriorhodopsin dataset were accurately predicted by all three programs.

There are other stability predictors, in addition to those mentioned above, that we did not test. Eris (http://eris.dokhlab.org) uses a physical force field in combination with atomic modeling and fast side-chain packing (Yin et al. 2007). The program is also designed to predict changes in backbone conformations caused by mutations by modeling backbone flexibility. Because the Eris website does not allow for batch submissions, we could not study its performance. iPTREE-STAB (http://210.60.98.17/IPTREEr/iptree.htm) uses a decision-tree method. The sequence-based method determines stabilizing and destabilizing mutations but uses only a seven-residue window, the mutation position in the middle. The service could not be accessed. Finally, although we attempted to assess the prediction accuracy of AUTO-MUTE, only 28 cases that had not been used to train the program could be retrieved from ProTherm, which was too small a number for a statistical analysis. Of the 28 cases, AUTO-MUTE correctly predicted 6 (21%).

Overall, we found SRide to be the least accurate predictor and that SCide and MUpro also performed poorly. The latter two predictors use machine-learning approaches that are dependent on the quality and quantity of the training dataset.

Mutations can introduce or relieve strain into the protein backbone. To properly estimate ΔΔG stability values, structural rearrangements that induce or release strain should be considered. Calculations of the $\Delta\Delta G$ values associated with strain are computationally possible using either molecular dynamics or Monte Carlo simulations but are also computationally very intense. The simpler methods, such as those that we

used, allow a large number of mutations to be surveyed and their effects on stabilities determined quickly but can not model protein dynamics.

Our analyses showed that the predicted ΔΔG values are distributed in a fashion similar to those of the experimental data. However, the mutant and mutated residue distributions are strongly biased in the stabilizing, destabilizing, and neutral categories. These biases may have arisen because the designs of the original experiments that produced the mutations were biased, e.g., consider the excessive number of alanine mutations retrieved from ProTherm.

Our ROC curves are quite similar to those found for a function-stability correlation study that used missense mutations (Bromberg and Rost 2009). The curves in Fig. 2 increase sharply until a tp value of 0.6 is reached, but then bend sharply, and continue to rise more slowly.

We found that the structural context of a residue strongly affected predictor performance. Disease-causing mutations have biased distributions in secondary structural elements (Khan and Vihinen 2007). Both the secondary structure type and the protein folding type had significant effects. There was also a clear difference between the prediction accuracies for buried and accessible residues. The structural context effect depended on the method used and influenced the values of the quality parameters differently. Conversely, the extent of volume or charge change upon mutation did not influence the prediction performances significantly.

In conclusion, at best, the methods predicted the changes in stability caused by mutations with only moderate accuracies. However, the number of false positives and false negatives returned by the programs was substantial. As so many factors affect protein stability, even small differences in the ΔΔG values between a wild-type and its

mutant can be significant. Molecular dynamics and Monte Carlo simulations provide more accurate results in general; however, characterization of mutational effects is still problematic even when these methods are used. Additionally, the computational power demands of these two methods are prohibitively great for the analysis of large datasets. For mutation effect investigations the tested methods have only limited applicability, and should thus be used preferably together with other prediction approaches. One way to improve the performance of predictors might be to use additional features.

Acknowledgments

Financial support from the Finnish Academy, the Sigrid Juselius Foundation, and the Medical Research Fund of Tampere University Hospital is gratefully acknowledged.

References

Abkevich VI, Gutin AM and Shakhnovich EI (1995): Impact of local and non-local interactions on thermodynamics and kinetics of protein folding. J Mol Biol 252:460-471.

Akasako A, Haruki M, Oobatake M and Kanaya S (1997): Conformational stabilities of Escherichia coli RNase HI variants with a series of amino acid substitutions at a cavity within the hydrophobic core. J Biol Chem 272:18686-18693.

Baltzer L and Nilsson J (2001): Emerging principles of de novo catalyst design. Curr Opin Biotechnol 12:355-360.

Bash PA, Singh UC, Langridge R and Kollman PA (1987): Free energy calculations by computer simulation. Science 236:564-568.

Beadle BM and Shoichet BK (2002): Structural bases of stability-function tradeoffs in enzymes. J Mol Biol 321:285-296.

Benedix A, Becker CM, de Groot BL, Caflisch A and Bockmann RA (2009): Predicting free energy changes using structural ensembles. Nat Methods 6:3-4.

Bloom JD, Meyer MM, Meinhold P, Otey CR, MacMillan D and Arnold FH (2005): Evolving strategies for enzyme engineering. Curr Opin Struct Biol 15:447-452.

Bloom JD, Raval A and Wilke CO (2007): Thermodynamics of neutral protein evolution. Genetics 175:255-266.

Bloom JD, Silberg JJ, Wilke CO, Drummond DA, Adami C and Arnold FH (2005): Thermodynamic prediction of protein neutrality. Proc Natl Acad Sci U S A 102:606-611.

Bolon DN, Voigt CA and Mayo SL (2002): De novo design of biocatalysts. Curr Opin Chem Biol 6:125-129.

Bromberg Y and Rost B (2009): Correlating protein function and stability through the analysis of single amino acid substitutions. BMC Bioinformatics 10 Suppl 8:S8.

Bross P, Corydon TJ, Andresen BS, Jorgensen MM, Bolund L and Gregersen N (1999): Protein misfolding and degradation in genetic diseases. Hum Mutat 14:186-198.

Butterfoss GL and Kuhlman B (2006): Computer-based design of novel protein structures. Annu Rev Biophys Biomol Struct 35:49-65.

Camps M, Herman A, Loh E and Loeb LA (2007): Genetic constraints on protein evolution. Crit Rev Biochem Mol Biol 42:313-326.

Capriotti E, Fariselli P and Casadio R (2004): A neural-network-based method for predicting protein stability changes upon single point mutations. Bioinformatics 20 Suppl 1:i63-68.

Capriotti E, Fariselli P and Casadio R (2005): I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res 33:W306-310.

Capriotti E, Fariselli P, Rossi I and Casadio R (2008): A three-state prediction of single point mutations on protein stability changes. BMC Bioinformatics 9 Suppl 2:S6.

Cheng J, Randall A and Baldi P (2006): Prediction of protein stability changes for single-site mutations using support vector machines. Proteins 62:1125-1132.

Counago R, Wilson CJ, Pena MI, Wittung-Stafshede P and Shamoo Y (2008): An adaptive mutation in adenylate kinase that increases organismal fitness is linked to stability-activity trade-offs. Protein Eng Des Sel 21:19-27.

Cunningham BC and Wells JA (1989): High-resolution epitope mapping of hGH-receptor interactions by alanine-scanning mutagenesis. Science 244:1081-1085.

Daniel RM, Dunn RV, Finney JL and Smith JC (2003): The role of dynamics in enzyme activity. Annu Rev Biophys Biomol Struct 32:69-92.

Dehouck Y, Grosfils A, Folch B, Gilis D, Bogaerts P and Rooman M (2009): Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. Bioinformatics 25:2537-2543.

DePristo MA, Weinreich DM and Hartl DL (2005): Missense meanderings in sequence space: a biophysical view of protein evolution. Nat Rev Genet 6:678-687.

Deutsch C and Krishnamoorthy B (2007): Four-body scoring function for mutagenesis. Bioinformatics 23:3009-3015.

Dosztanyi Z, Fiser A and Simon I (1997): Stabilization centers in proteins: identification, characterization and predictions. J Mol Biol 272:597-612.

Dosztanyi Z, Magyar C, Tusnady G and Simon I (2003): SCide: identification of stabilization centers in proteins. Bioinformatics 19:899-900.

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

Fersht AR (1999). Structure and Mechanism in Protein Science: a Guide to Enzyme Catalysis and Protein Folding. New York, W. H. Freeman.

Fields PA (2001): Review: Protein function at thermal extremes: balancing stability and flexibility. Comp Biochem Physiol A Mol Integr Physiol 129:417-431.

Fink AL (2005): Natively unfolded proteins. Curr Opin Struct Biol 15:35-41.

Garcia C, Nishimura C, Cavagnero S, Dyson HJ and Wright PE (2000): Changes in the apomyoglobin folding pathway caused by mutation of the distal histidine residue. Biochemistry 39:11227-11237.

Gilis D and Rooman M (1997): Predicting protein stability changes upon mutation using database-derived potentials: solvent accessibility determines the importance of local versus non-local interactions along the sequence. J Mol Biol 272:276-290.

Gilis D and Rooman M (2000): PoPMuSiC, an algorithm for predicting protein mutant stability changes: application to prion proteins. Protein Eng 13:849-856.

Greenstein JP and Winitz M (1961). Chemistry of the Amino Acids. New York, John Wiley & Sons.

Gromiha MM and Selvaraj S (2004): Inter-residue interactions in protein folding and stability. Prog Biophys Mol Biol 86:235-277.

Guerois R, Nielsen JE and Serrano L (2002): Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. J Mol Biol 320:369-387.

Hsu YH, Johnson DA and Traugh JA (2008): Analysis of conformational changes during activation of protein kinase Pak2 by amide hydrogen/deuterium exchange. J Biol Chem 283:36397-36405.

Khan S and Vihinen M (2007): Spectrum of disease-causing mutations in protein secondary structures. BMC Struct Biol 7:56.

Khatun J, Khare SD and Dokholyan NV (2004): Can contact potentials reliably predict stability of proteins? J Mol Biol 336:1223-1238.

Kidokoro S, Miki Y, Endo K, Wada A, Nagao H, Miyake T, Aoyama A, Yoneya T, Kai K and Ooe S (1995): Remarkable activity enhancement of thermolysin mutants. FEBS Lett 367:73-76.

Kumar MD, Bava KA, Gromiha MM, Prabakaran P, Kitajima K, Uedaira H and Sarai A (2006): ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. Nucleic Acids Res 34:D204-206.

Lehmann M and Wyss M (2001): Engineering proteins for thermostability: the use of sequence alignments versus rational design and directed evolution. Curr Opin Biotechnol 12:371-375.

Lonquety M, Lacroix Z and Chomilier J (2008). Evaluation of the stability of folding nucleus upon mutation. Pattern Recognition in Bioinformatics. M. Chetty, S. Ahmad and A. Ngom. Heidelberg, Springer. 5265.

Magyar C, Gromiha MM, Pujadas G, Tusnady GE and Simon I (2005): SRide: a server for identifying stabilizing residues in proteins. Nucleic Acids Res 33:W303-305.

Markert Y, Koditz J, Mansfeld J, Arnold U and Ulbrich-Hofmann R (2001): Increased proteolytic resistance of ribonuclease A by protein engineering. Protein Eng 14:791-796.

Masso M and Vaisman, II (2008): Accurate prediction of stability changes in protein mutants by combining machine learning with structure based computational mutagenesis. Bioinformatics 24:2002-2009.

Meiering EM, Serrano L and Fersht AR (1992): Effect of active site residues in barnase on activity and stability. J Mol Biol 225:585-589.

Mohamed AJ, Yu L, Bäckesjö CM, Vargas L, Faryal R, Aints A, Christensson B, Berglöf A, Vihinen M, Nore BF and Smith CIE (2009): Bruton's tyrosine kinase (Btk): function, regulation, and transformation with special emphasis on the PH domain. Immunol Rev 228:58-73.

Mukaiyama A, Haruki M, Ota M, Koga Y, Takano K and Kanaya S (2006): A hyperthermophilic protein acquires function at the cost of stability. Biochemistry 45:12673-12679.

Muller CW, Schlauderer GJ, Reinstein J and Schulz GE (1996): Adenylate kinase motions during catalysis: an energetic counterweight balancing substrate binding. Structure 4:147-156.

Nagatani RA, Gonzalez A, Shoichet BK, Brinen LS and Babbitt PC (2007): Stability for function trade-offs in the enolase superfamily "catalytic module". Biochemistry 46:6688-6695.

Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB and Thornton JM (1997): CATH--a hierarchic classification of protein domain structures. Structure 5:1093-1108.

Pace CN (1990): Conformational stability of globular proteins. Trends Biochem Sci 15:14-17.

Pal C, Papp B and Lercher MJ (2006): An integrated view of protein evolution. Nat Rev Genet 7:337-348.

Parthiban V, Gromiha MM and Schomburg D (2006): CUPSAT: prediction of protein stability upon point mutations. Nucleic Acids Res 34:W239-242.

Petsko GA (2001): Structural basis of thermostability in hyperthermophilic proteins, or "there's more than one way to skin a cat". Methods Enzymol 334:469-478.

Pitera JW and Kollman PA (2000): Exhaustive mutagenesis in silico: multicoordinate free energy calculations on proteins and peptides. Proteins 41:385-397.

Poelwijk FJ, Kiviet DJ, Weinreich DM and Tans SJ (2007): Empirical fitness landscapes reveal accessible evolutionary paths. Nature 445:383-386.

Pokala N and Handel TM (2005): 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 347:203-227.

Ponnuswamy PK and Gromiha MM (1994): On the conformational stability of oligonucleotide duplexes and tRNA molecules. J Theor Biol 169:419-432.

Pontius J, Richelle J and Wodak SJ (1996): Deviations from standard atomic volumes as a quality measure for protein crystal structures. J Mol Biol 264:121-136.

Potapov V, Cohen M and Schreiber G (2009): Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details. Protein Eng Des Sel 22:553-560.

Prevost M, Wodak SJ, Tidor B and Karplus M (1991): Contribution of the hydrophobic effect to protein stability: analysis based on simulations of the Ile-96----Ala mutation in barnase. Proc Natl Acad Sci U S A 88:10880-10884.

Sawano M, Yamamoto H, Ogasahara K, Kidokoro S, Katoh S, Ohnuma T, Katoh E, Yokoyama S and Yutani K (2008): Thermodynamic basis for the stabilities of three CutA1s from Pyrococcus horikoshii,Thermus thermophilus, and Oryza sativa, with unusually high denaturation temperatures. Biochemistry 47:721-730.

Schreiber G, Buckle AM and Fersht AR (1994): Stability and function: two constraints in the evolution of barstar and other proteins. Structure 2:945-951.

Shen B, Bai J and Vihinen M (2008): Physicochemical feature-based classification of amino acid mutations. Protein Eng Des Sel 21:37-44.

Shoichet BK, Baase WA, Kuroki R and Matthews BW (1995): A relationship between protein stability and protein function. Proc Natl Acad Sci U S A 92:452-456.

Somero GN (1995): Proteins and temperature. Annu Rev Physiol 57:43-68.

Steward RE, MacArthur MW, Laskowski RA and Thornton JM (2003): Molecular basis of inherited diseases: a structural perspective. Trends Genet 19:505-513.

Sunyaev S, Lathe W, 3rd and Bork P (2001): Integration of genome data and protein structures: prediction of protein folds, protein interactions and "molecular phenotypes" of single nucleotide polymorphisms. Curr Opin Struct Biol 11:125-130.

Tastan O, Yu E, Ganapathiraju M, Aref A, Rader AJ and Klein-Seetharaman J (2007): Comparison of stability predictions and simulated unfolding of rhodopsin structures. Photochem Photobiol 83:351-362.

Thusberg J and Vihinen M (2009): Pathogenic or not? And if so, then how? Studying the effects of missense mutations using bioinformatics methods. Hum Mutat 30:703-714.

van den Burg B and Eijsink VG (2002): Selection of mutations for increased protein stability. Curr Opin Biotechnol 13:333-337.

Wang G and Dunbrack RL, Jr. (2003): PISCES: a protein sequence culling server. Bioinformatics 19:1589-1591.

Wang Z and Moult J (2001): SNPs, protein structure, and disease. Hum Mutat 17:263-270.

Vihinen M (1987): Relationship of protein flexibility to thermostability. Protein Eng 1:477-480.

Villegas V, Viguera AR, Aviles FX and Serrano L (1996): Stabilization of proteins by rational design of α-helix stability using helix/coil transition theory. Fold Des 1:29-34.

Wolf-Watz M, Thai V, Henzler-Wildman K, Hadjipavlou G, Eisenmesser EZ and Kern D (2004): Linkage between dynamics and catalysis in a thermophilic-mesophilic enzyme pair. Nat Struct Mol Biol 11:945-949.

Yin S, Ding F and Dokholyan NV (2007): Modeling backbone flexibility improves protein stability estimation. Structure 15:1567-1576.

Yue P, Li Z and Moult J (2005): Loss of protein structure stability as a major causative factor in monogenic disease. J Mol Biol 353:459-473.

Zhi W, Srere PA and Evans CT (1991): Conformational stability of pig citrate synthase and some active-site mutants. Biochemistry 30:9281-9286.

Zhou H and Zhou Y (2002): Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci 11:2714-2726.

Zhou H and Zhou Y (2004): Quantifying the effect of burial of amino acid residues on protein stability. Proteins 54:315-322.

Figure Legends

Figure 1. A) Distributions of predicted and experimental ΔΔG values. The predictors used were I-Mutant2.0 (red), Dmutant (green), CUPSAT (blue), FoldX (grey), and the experimental ΔΔG values are shown in black. B) Receiver operating characteristics curves diagramming the performances of FoldX, I-Mutant2.0, Dmutant and CUPSAT with the values for AUC ± SE derived from the areas under the curves. Color coding for the individual predictors is shown in the figure.

Figure 2. The values of the four quality parameters, accuracy, specificity, sensitivity, and Matthew's correlation coefficient for the secondary structures, the CATH classifications, and the accessible surface areas. A) Secondary structures: α-helices (red), β-strands (blue), coils (yellow), and turns (green). B) Protein structure types: mainly α-helical (red), mainly β-stranded (blue), α/β structures (green), and aperiodic structures (yellow). C) Accessible surface areas: exposed residues (blue, ASA ≥ 25%) and buried residues (red, ASA ≤ 10%). Color coding for the classifications is shown in the figure.

Figure 3. Number of stability predictors that returned predictions that agreed with the experimental values. Black bars do not include the results of the stability-center programs (SCide, SRide and Scpred). The grey bars include the results of all of the programs. The signs of the tp, fp, tn, and fn values were taken into account.

Table 1. Performance of stability predictors

All cases

Parameters	CUPSAT	Dmutant	FoldX	I-Mutant2.0	I-Mutant3.0 (sequence)	I-Mutant3.0 (structure)	MUpro	MultiMutate	SCide	SRide	Scpred
tp	249	576	629	72	35	34	71	620	197	33	402
fp	123	238	321	53	38	23	70	414	122	28	238
tn	53	365	244	19	24	39	0	206	465	548	393
fn	111	535	347	30	18	19	25	517	862	980	751
Total[a]	536	1714	1541	174	115	115	166	1757	1646	1589	1784
Accuracy[b]	0.50	0.56	0.54	0.48	0.52	0.64	0.37	0.44	0.49	0.49	0.49
Specificity[b]	0.50	0.57	0.53	0.49	0.52	0.63	0.43	0.45	0.47	0.40	0.48
Sensitivity[b]	0.69	0.52	0.64	0.71	0.66	0.64	0.74	0.55	0.19	0.03	0.35
MCC[b]	-0.01	0.12	0.08	-0.03	0.05	0.27	-0.39	-0.13	-0.03	-0.04	-0.03

Stability increasing cases

Parameters	CUPSAT	Dmutant	FoldX	I-Mutant2.0	MUpro	MultiMutate
tp	25	91	86	8	8	91
fp	45	131	134	7	15	193
tn	131	472	431	65	55	427
fn	33	123	125	15	17	128
Total[a]	234	817	776	95	95	839
Accuracy[b]	0.74	0.60	0.59	0.63	0.55	0.55
Specificity[b]	0.63	0.66	0.63	0.78	0.60	0.57
Sensitivity[b]	0.43	0.43	0.41	0.35	0.32	0.42
MCC[b]	0.35	0.22	0.18	0.30	0.12	0.11

Stability decreasing cases

Parameters	CUPSAT	Dmutant	FoldX	I-Mutant2.0	I-Mutant3.0 (sequence)	I-Mutant3.0 (structure)	MUpro	MultiMutate
tp	224	485	543	64	35	34	63	529
fp	78	107	187	46	36	20	55	221
tn	98	496	378	26	26	42	15	399

fn	78	412	222	15	13	14	8	389
Total[a]	478	1500	1330	151	110	110	141	1538
Accuracy[b]	0.65	0.68	0.69	0.59	0.57	0.69	0.55	0.61
Specificity[b]	0.63	0.75	0.68	0.56	0.56	0.69	0.53	0.62
Sensitivity[b]	0.74	0.54	0.71	0.81	0.73	0.71	0.89	0.58
MCC[b]	0.30	0.38	0.38	0.19	0.16	0.39	0.14	0.22

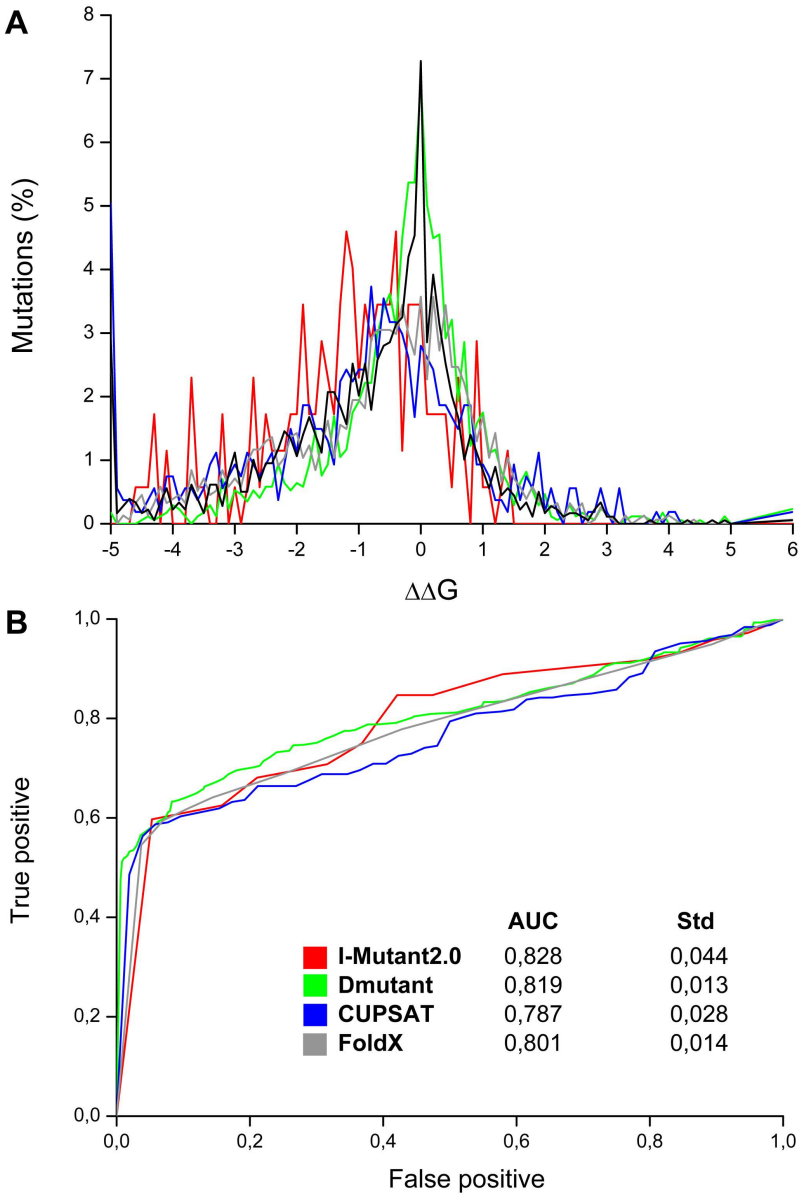
[a]Total number of cases used by the given program.
[b]Accuracy, specificity, sensitivity and MCC are calculated from normalized numbers.

Table 2. Pairwise prediction correlations. Upper table: The number of cases shared by two programs, reported as a percentage (upper right triangle). The number of cases predicted correctly, reported as a percentage (lower left triangle). Middle table: The absolute number of cases shared by two programs (upper right triangle). The percentage of correctly predicted cases (lower left triangle). Bottom table: Pairwise correlation

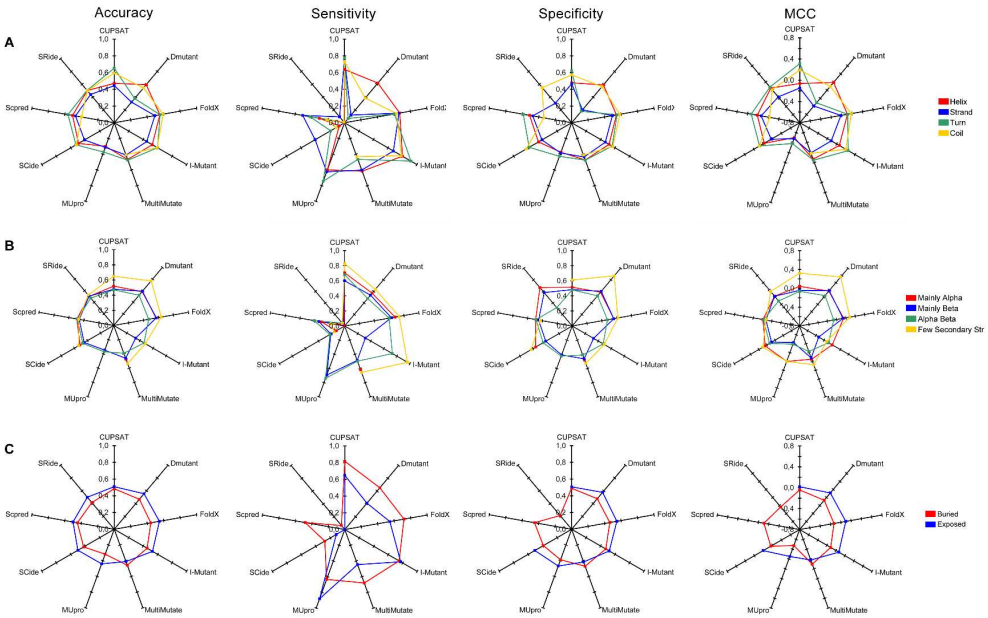
	CUPSAT	Dmutant	FoldX	I-Mutant2.0	I-Mutant3.0 structure	I-Mutant3.0 sequence	MultiMutate	MUpro	SCide	Scpred	SRide
CUPSAT		29.4	21.4	1.0	0.1	0.1	29.5	2.0	26.1	30.0	23.0
Dmutant	8.1		82.5	9.6	6.3	6.3	94.6	9.1	90.4	96.1	87.7
FoldX	8.1	31.5		9.8	6.4	6.4	84.9	9.3	78.7	86.2	75.3
I-Mutant2.0	0.2	3.6	3.5		6.4	6.4	9.5	7.5	9.0	9.8	9.0
I-Mutant3.0 structure	0.0	3.2	3.0	2.1		6.4	6.4	5.9	6.4	6.4	6.4
I-Mutant3.0 sequence	0.0	2.6	2.4	2.4	2.7		6.4	5.9	6.4	6.4	6.4
MultiMutate	7.8	30.7	27.2	3.3	2.2	2.0		9.2	90.8	98.5	87.7
MUpro	0.7	2.9	2.9	2.0	1.5	1.7	2.6		8.5	9.3	8.5
Scide	4.5	22.5	17.6	1.4	2.6	1.8	16.2	1.2		92.3	87.8
Scpred	7.5	26.5	22.9	2.3	2.9	2.3	21.6	2.6	24.6		89.1
Sride	2.6	19.9	13.5	1.2	2.4	1.6	11.5	0.4	26.2	19.8	

	CUPSAT	Dmutant	FoldX	I-Mutant2.0	I-Mutant3.0 structure	I-Mutant3.0 sequence	MultiMutate	MUpro	SCide	Scpred	SRide
CUPSAT		524	381	18	1	1	527	35	465	536	411
Dmutant	27		1471	171	113	113	1688	162	1613	1714	1565
FoldX	38	38		174	115	115	1514	166	1404	1538	1344
I-Mutant2.0	22	38	36		114	114	169	134	160	174	161
I-Mutant3.0 structure	0	50	46	33		115	114	106	115	115	115
I-Mutant3.0 sequence	0	42	37	37	43		114	106	115	115	115
MultiMutate	26	32	32	35	34	31		164	1620	1757	1564
MUpro	34	32	31	27	25	28	28		152	166	152
Scide	17	25	22	16	40	28	18	14		1646	1566
Scpred	25	28	27	24	45	36	22	28	27		1589
Sride	11	23	18	14	37	25	13	5	30	22	

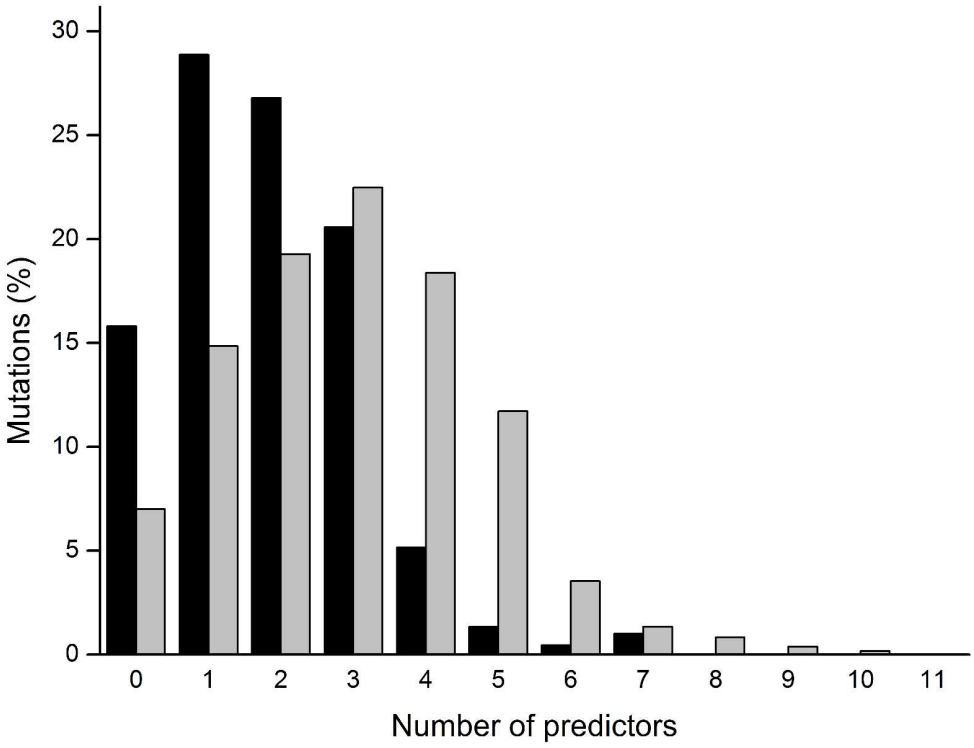
	CUPSAT	Dmutant	FoldX	I-Mutant2.0	I-Mutant3.0 structure	I-Mutant3.0 sequence	MultiMutate	MUpro	SCide	Scpred	SRide
CUPSAT											
Dmutant	0.04										
FoldX	0.28	0.28									
I-Mutant2.0	0.16	0.18	0.24								
I-Mutant3.0 structure	-	0.38	0.38	0.17							
I-Mutant3.0 sequence	-	0.33	0.27	0.53	0.42						
MultiMutate	0.15	0.25	0.20	0.26	0.04	0.16					
MUpro	0.54	0.09	0.29	0.37	0.02	0.33	0.23				
Scide	-0.14	0.10	-0.03	-0.26	0.24	0.01	-0.05	-0.30			
Scpred	-0.07	0.12	0.06	0.07	0.44	0.30	0.04	0.22	0.35		
Sride	-0.28	0.10	-0.15	-0.37	0.07	-0.12	-0.18	-0.65	0.64	0.22	



Distributions of predicted and experimental ΔΔG values. The predictors used were I-Mutant2.0 (red), Dmutant (green), CUPSAT (blue), FoldX (grey), and the experimental ΔΔG values are shown in black. B) Receiver operating characteristics curves diagramming the performances of FoldX, I-Mutant2.0, Dmutant and CUPSAT with the values for AUC ± SE derived from the areas under the curves. Color coding for the individual predictors is shown in the figure.
181x261mm (300 x 300 DPI)



The values of the four quality parameters, accuracy, specificity, sensitivity, and Matthew's correlation coefficient for the secondary structures, the CATH classifications, and the accessible surface areas. A) Secondary structures: α-helices (red), β-strands (blue), coils (yellow), and turns (green). B) Protein structure types: mainly α-helical (red), mainly β-stranded (blue), α/β structures (green), and aperiodic structures (yellow). C) Accessible surface areas: exposed residues (blue, ASA ≥ 25%) and buried residues (red, ASA ≤ 10%). Color coding for the classifications is shown in the figure.
297x187mm (300 x 300 DPI)



Number of stability predictors that returned predictions that agreed with the experimental values. Black bars do not include the results of the stability-center programs (SCide, SRide and Scpred). The grey bars include the results of all of the programs. The signs of the tp, fp, tn, and fn values were taken into account.
123113x95250mm (1 x 1 DPI)

Supplemental Table 1. Amino acid distributions. A) Mutated (original) and B) mutant amino acids

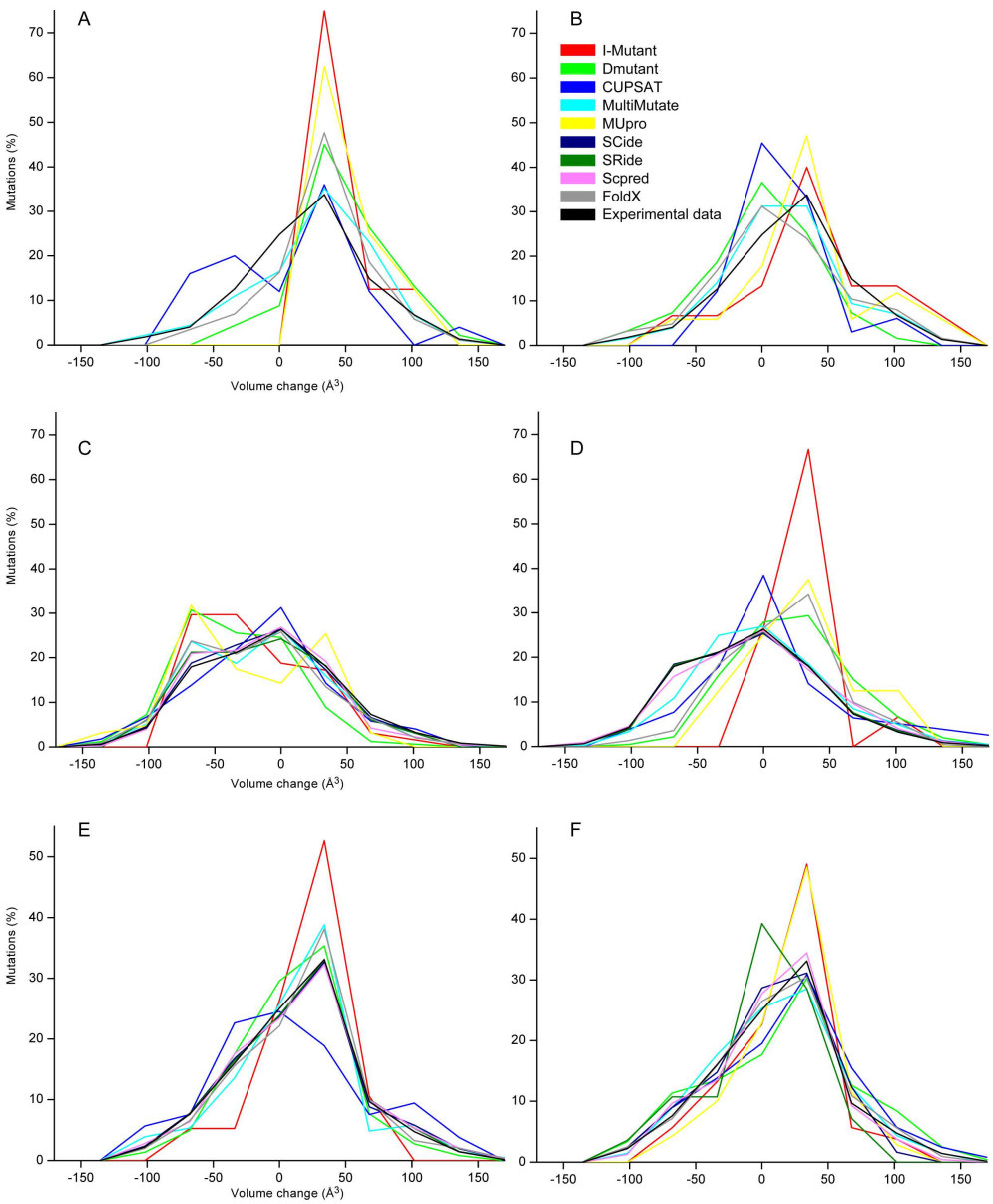
A)	Stabilizing				Destabilizing				Neutral				
Amino acid	Observed	Expected	χ^2	P value	Observed	Expected	χ^2	P value	Observed	Expected	χ^2	P value	Total
A	18	19	0.03	8.64E-01	70	79	0.94	3.32E-01	53	53	0	9.70E-01	141
C	4	3	0.17	6.76E-01	27	14	**13.15*** **	2.88E-04	5	9	1.94	1.64E-01	36
D	36	14	**34.13*** **	5.15E-09	46	59	2.88	8.96E-02	57	40	**7.21** **	7.26E-03	139
E	14	14	0	9.90E-01	33	59	*11.12** **	8.53E-04	69	40	**21.71*** **	3.17E-06	116
F	5	8	1.17	2.79E-01	33	34	0.02	8.83E-01	13	23	*4.31**	3.78E-02	51
G	10	19	*4.31**	3.78E-02	40	80	*19.98** **	7.84E-06	29	54	*11.72** **	6.20E-04	79
H	12	5	**11.55*** **	6.79E-04	18	20	0.12	7.25E-01	18	13	1.7	1.92E-01	48
I	10	11	0.12	7.29E-01	114	47	**96.52*** **	8.81E-23	37	32	0.88	3.48E-01	161
K	12	17	1.29	2.56E-01	28	70	*25.01** **	5.71E-07	47	47	0	9.66E-01	87
L	7	18	*6.61**	1.01E-02	96	75	**5.91***	1.50E-02	28	51	*10.23***	1.38E-03	131
M	6	5	0.47	4.92E-01	24	19	1.3	2.54E-01	16	13	0.75	3.87E-01	46
N	14	11	1.01	3.14E-01	29	45	*5.63**	1.77E-02	25	30	0.97	3.25E-01	68
P	3	9	*4.11**	4.26E-02	35	38	0.28	5.97E-01	6	26	*15.33** **	9.04E-05	44
Q	9	9	0	9.81E-01	13	38	*16.49** **	4.90E-05	23	26	0.3	5.83E-01	45
R	5	9	1.62	2.04E-01	19	37	*8.57***	3.41E-03	25	25	0	9.85E-01	49
S	13	13	0	9.58E-01	26	55	*15.54** **	8.09E-05	38	37	0.01	9.34E-01	77
T	19	13	3.19	7.40E-02	67	53	3.68	5.51E-02	46	36	2.81	9.35E-02	132
V	13	15	0.18	6.74E-01	149	61	**125.72*** **	3.55E-29	60	42	**8.23***	4.11E-03	222
W	0	4	3.76	5.26E-02	15	16	0.04	8.50E-01	7	11	1.27	2.61E-01	22
Y	12	8	1.87	1.72E-01	49	34	**6.61***	1.02E-02	29	23	1.54	2.15E-01	90
total	222	222			931	931			631	631			1784

B)	Stabilizing				Destabilizing				Neutral				
Amino acid	Observed	Expected	χ^2	P value	Observed	Expected	χ^2	P value	Observed	Expected	χ^2	P value	Total
A	25	14	**9.58****	1.97E-03	307	57	**1096.49*****	1.91E-240	138	39	**255.58*****	1.57E-57	470
C	12	8	2.09	1.48E-01	25	33	2.05	1.53E-01	12	23	*4.93**	2.65E-02	49
D	5	9	1.82	1.77E-01	25	38	*4.45**	3.50E-02	24	26	0.12	7.29E-01	54
E	8	8	0	9.80E-01	15	33	*10.02***	1.55E-03	34	23	**5.83***	1.57E-02	57
F	12	9	0.95	3.29E-01	47	38	2.13	1.44E-01	35	26	3.32	6.85E-02	94
G	14	13	0.07	7.87E-01	99	55	**36.05*****	1.92E-09	23	37	*5.31**	2.12E-02	136
H	9	9	0	9.84E-01	23	38	*5.92**	1.50E-02	17	26	2.98	8.45E-02	49
I	19	12	**4.25***	3.93E-02	26	50	*11.43****	7.23E-04	24	34	2.84	9.18E-02	69
K	9	8	0.14	7.04E-01	25	33	2.05	1.53E-01	39	23	**12.03*****	5.24E-04	73
L	18	19	0.03	8.73E-01	33	78	*26.27****	2.97E-07	32	53	*8.40***	3.76E-03	83
M	10	5	**4.72***	2.99E-02	31	21	**4.33***	3.74E-02	22	14	3.9	4.84E-02	63
N	9	9	0	9.84E-01	30	38	1.68	1.94E-01	23	26	0.29	5.87E-01	62
P	4	14	*6.77***	9.28E-03	20	57	*24.02****	9.55E-07	14	39	*15.71****	7.40E-05	38
Q	5	8	1.08	2.98E-01	21	33	*4.51**	3.36E-02	34	23	**5.83***	1.57E-02	60
R	11	19	3.54	5.99E-02	9	81	*63.75****	1.41E-15	22	55	*19.57****	9.68E-06	42
S	13	21	3.02	8.23E-02	51	88	*15.47****	8.37E-05	36	60	*9.32***	2.27E-03	100
T	4	14	*6.77***	9.28E-03	55	57	0.07	7.91E-01	20	39	*8.99***	2.72E-03	79
V	22	14	**5.20***	2.26E-02	64	57	0.86	3.54E-01	50	39	3.34	6.74E-02	136
W	5	4	0.27	6.03E-01	9	17	3.5	6.15E-02	10	11	0.14	7.06E-01	24
Y	8	7	0.21	6.44E-01	16	29	*5.48**	1.92E-02	22	19	0.37	5.41E-01	46
total	222	222			931	931			631	631			1784

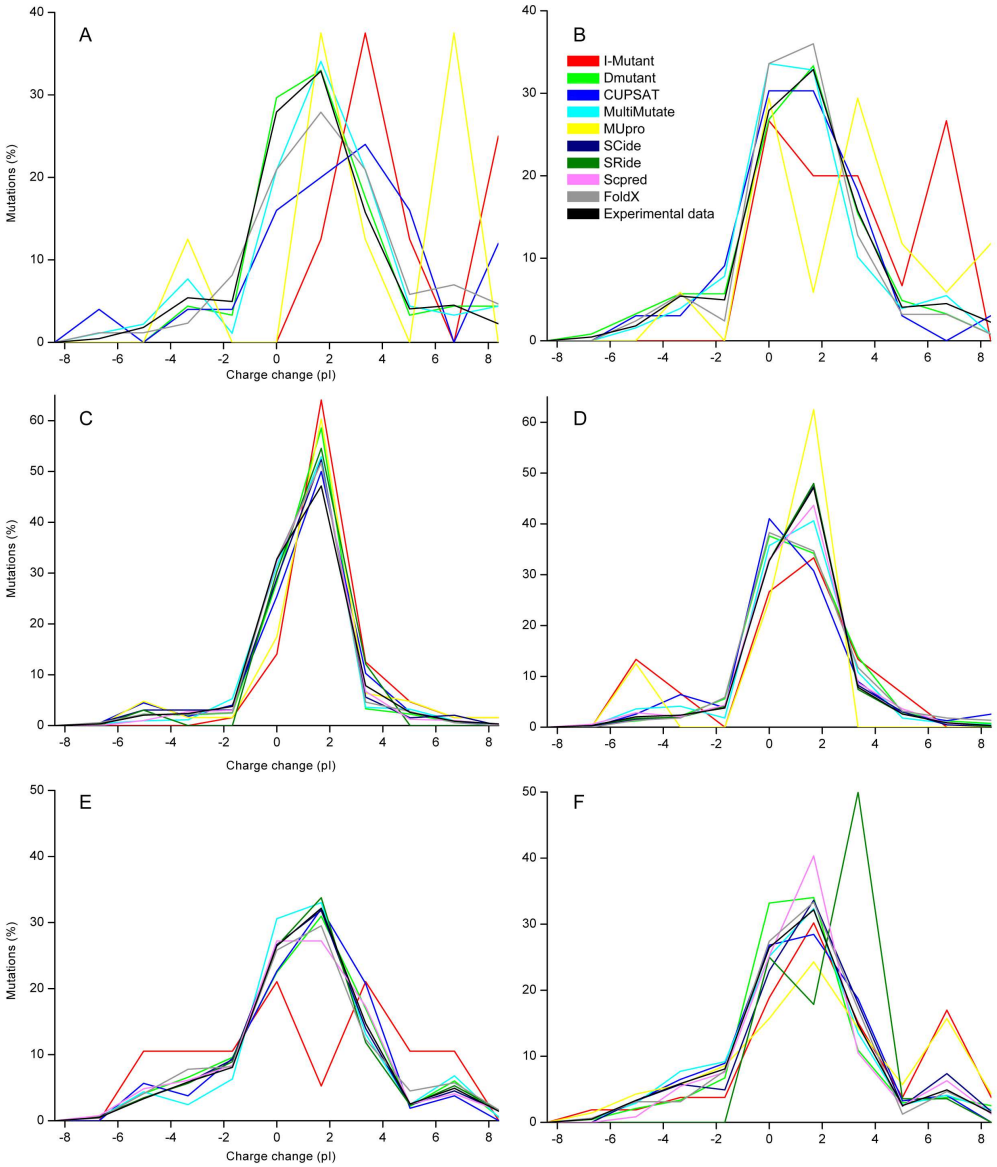
[a]The italicized χ^2 values identify underrepresented residues and the values in bold identify overrepresented residues in comparison with random distributions derived from theoretical usage frequencies. Significance levels are * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$

Supplementary Figure 1. The performances of the predictors as a function of the volume change resulting from mutation. A) Stabilizing true positives, B) stabilizing false negatives, C) destabilizing true positives, D) destabilizing false negatives, E) neutral true negatives, and F) neutral false positives. Color coding for the individual programs and the complete dataset is shown in the figure.

Supplementary Figure 2. The performances of the predictors as a function of the charge change resulting from mutation. A) Stabilizing true positives, B) stabilizing false negatives, C) destabilizing true positives, D) destabilizing false negatives, E) neutral true negatives, and F) neutral false positives. Color coding for the individual programs and the complete dataset is shown in the figure.



The performances of the predictors as a function of the volume change resulting from mutation. A) Stabilizing true positives, B) stabilizing false negatives, C) destabilizing true positives, D) destabilizing false negatives, E) neutral true negatives, and F) neutral false positives. Color coding for the individual programs and the complete dataset is shown in the figure.
209x252mm (300 x 300 DPI)



The performances of the predictors as a function of the charge change resulting from mutation. A) Stabilizing true positives, B) stabilizing false negatives, C) destabilizing true positives, D) destabilizing false negatives, E) neutral true negatives, and F) neutral false positives. Color coding for the individual programs and the complete dataset is shown in the figure.
209x245mm (300 x 300 DPI)