Prediction of protein stability upon point mutations

M.M. Gromiha1
Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), AIST Tokyo Waterfront Bio-IT Research Building, 2–42 Aomi, Koto-ku, Tokyo 135-0064, Japan

Abstract
Prediction of protein stability upon amino acid substitution is a challenging problem and it will be helpful for designing stable mutants. We have developed a thermodynamic database for proteins and mutants (ProTherm), which has more than 20 000 thermodynamic data along with sequence and structure information, experimental conditions and literature information. It is freely accessible at http://gibk26.bse.kyutech.ac.jp/jouhou/protherm/protherm.html. Utilizing the database, we have analysed the relationship between amino acid properties and protein stability and developed different methods, such as average assignment method, distance and torsion potentials and decision tree models to discriminate the stabilizing and destabilizing mutants, and to predict the stability change upon mutation. Our method could distinguish the stabilizing and destabilizing mutants with an accuracy of 82 and 85% respectively from amino acid sequence and protein three-dimensional structure. We obtained the correlation of 0.70 and 0.87, between the experimental and predicted stability changes upon mutations, from sequence and structure respectively. Furthermore, we have developed different web servers for discrimination and prediction and they are freely accessible at http://bioinformatics.myweb.hinet.net/iptree.htm and http://cupsat.tu-bs.de/.

Introduction
Protein structures are stabilized with various non-covalent interactions such as hydrophobic, electrostatic, van der Waals and hydrogen bonds [1–3]. The importance of such interactions for protein stability has been revealed by site-directed mutagenesis experiments [4–8]. These experimental data have been collected together in the form of a book [9] and an electronically accessible database [10], which are valuable resources for the analysis and prediction of protein stability.

Several methods have been proposed for understanding the factors influencing the stability of protein mutants and predicting the stability of protein upon single mutations. These studies include the development of distance and torsion potentials [11,12], structural environment-dependent amino acid substitution and propensity tables [13], relationship between amino acid properties and protein stability [14–17], empirical energy functions [18,19], stability scale [20], contact potentials [21], support vector machines [23,24], decision trees [25], relative importance of secondary structure and solvent accessibility [26], average assignment [27] and Bayesian networks [28]. Furthermore, several web servers have been developed for predicting protein stability based on structural information as well as from amino acid sequence.

The present review is mainly focused on three aspects: (i) development of a thermodynamic database for proteins and mutants, (ii) understanding the factors influencing protein mutant stability and (iii) prediction of protein stability upon mutation.

ProTherm, a thermodynamic database for proteins and mutants
ProTherm is a large collection of thermodynamic data on protein stability that has the following information [10,29].

Sequence and structure information
Name, source, length and molecular mass of the protein, codes for PIR (protein information resource) [30], Swiss-Prot [31] and PDB [32], EC number [33], mutation details [wild-type and mutant residue names, residue number and location of the mutant based on secondary structure and solvent accessibility, ASA (accessible surface area)] and number of transition states. The secondary structure and ASA have been computed with DSSP [34] and ASC [35] respectively. For each mutant, information about the surrounding residues is given for different radii (e.g. 0–4 and 4–8 Å; 1 Å = 0.1 nm).

Thermodynamic data obtained from denaturant denaturation experiments
Unfolding Gibbs free energy change (ΔG_H2O), ΔΔG_H2O [ΔG_H2O(mutant–wild)], midpoint (Cm), slope (m) and reversibility of denaturation.

Thermodynamic data obtained from thermal denaturation experiments
Unfolding Gibbs free energy change (ΔG), ΔΔG [ΔG(mutant–wild)], transition temperature (Tm), ΔTm
Factors influencing the stability of protein mutants

We have systematically analysed the relationship between amino acid properties and protein stability upon amino acid substitutions. We considered a set of 49 diverse amino acid properties, which fall into various categories, such as physical, chemical, energetic and conformational. These properties have been used for understanding/predicting the folding rates of proteins [36], stability of protein mutants and thermophilic proteins [28,37] etc. The numerical values for the selected 49 physicochemical, energetic and conformational properties of the 20 amino acid residues and their brief explanations are available at http://www.cbrc.jp/~gromiha/fold.rate/property.html.

We computed the mutation-induced changes in property values, \( \Delta P(i) \), using the equation [14],

\[
\Delta P(i) = P_{\text{mut}}(i) - P_{\text{wild}}(i),
\]

where \( P_{\text{mut}}(i) \) and \( P_{\text{wild}}(i) \) are the normalized property value of the \( i \)th mutant and wild-type residue respectively; \( i \) varies from 1 to \( N \), where \( N \) is the total number of mutants.

The computed \( \Delta P \) values were related with \( \Delta T_m \), \( \Delta G \) or \( \Delta G^{\text{H}_2\text{O}} \) by using correlation coefficient.

We have also analysed the influence of neighbouring residues of the mutant residue in the amino acid sequence

\[
\Delta P_{\text{seq}}(i) = \sum_{j=i-k}^{i+k} P_j - P_{\text{mut}}(i),
\]

where \( k \) is the window length, and surrounding residues that are close in protein three-dimensional structure

\[
\Delta P_{\text{str}}(i) = \sum n_j \cdot P_j - P_{\text{mut}}(i),
\]

where \( n_j \) is the total number of type \( j \) residues surrounding the \( i \)th residue of the protein within the sphere of radius 8 Å [38,39], and \( P_j \) is the property value of the type \( j \) residue.

We observed that the protein mutant stability strongly depend on secondary structure and location of residues based on ASA. In buried mutations, the properties reflecting hydrophobicity showed a strong correlation with stability, indicating the direct relationship between hydrophobicity and stability [14,15]. Furthermore, the inclusion of neighbouring and surrounding residues did not show any significant improvement in the correlation between amino acid properties and protein stability. This might be due to the hydrophobic environment of the mutant site, which is surrounded mainly by hydrophobic residues, and non-specific interactions dominate in the interior of proteins.

In partially buried and exposed mutations, the whole set of data did not show significant correlation with any of the properties. However, the classification of data based on the mutations in helical, strand, coil and turn regions and hydrophobic and hydrogen bonds improved the correlation between amino acid properties and protein stability significantly. Furthermore, the inclusion of neighbouring and surrounding residues remarkably improved the correlation in all the subgroups of mutations. This result indicates that the information from nearby polar/charged residues and/or the residues that are close in space is important for the stability of partially buried and exposed mutations. Detailed analysis showed that more than 50% of the neighbouring/surrounding residues are polar and charged, and hence the stability of partially buried/exposed mutations is influenced by hydrophobic, hydrogen-bonding and other polar interactions [14,16,17].

Prediction of protein mutant stability

We have proposed different methods for predicting protein mutant stability: (i) the average assignment method, which can be used as a baseline for prediction methods, (ii) predictions based on structural information and (iii) just from amino acid sequence.

The average assignment method

In the average assignment method, the dataset has been classified into 380 possible amino acid substitutions (20 amino acids and 19 substitutions). Considering the Ala → Cys mutant, we calculated the average stability of all Ala → Cys mutants and assigned the same for this specific mutant. This calculation was repeated for all the 380 pairs and assigned the stability values. The assigned \( \Delta T_m \) (\( \Delta G \) and \( \Delta G^{\text{H}_2\text{O}} \)) values are compared with experimental stability [27].

We observed that our method could distinguish the stabilizing and destabilizing mutants to an accuracy of 70–80% at different measures of stability. Furthermore, we have classified the mutants based on secondary structure and solvent accessibility and observed that the classification significantly improved the accuracy of prediction. The classification of mutants based on helix, strand and coil distinguished the
stabilizing/destabilizing mutants at an average accuracy of 82% and the correlation was 0.56; information about the location of residues at the interior, partially buried and surface of a protein correctly identified the stabilizing/destabilizing residues at an average accuracy of 81% and the correlation was 0.59. The nine subclassifications based on three secondary structures and solvent accessibilities improved the accuracy of assigning stabilizing/destabilizing mutants to an accuracy of 84% and the correlation was 0.60.
of 84–89% for the three datasets. The limitation of the method is the insufficient amount of data for several pairs of mutants.

**Prediction of protein stability using structural information**

We have analysed protein stability upon point mutations using distance-dependent residue pair potentials and torsion potentials [12]. These potentials mainly represent the effect of residues that are close in protein structures and nearby residues in protein sequences respectively. We have combined the potentials and developed a statistical model for predicting protein mutant stability. Our method showed the maximum correlation of 0.87 in a set of 1538 mutants with a standard error of 0.71 kcal/mol (1 cal ≈ 4.184 J) between predicted and measured ΔΔG values and a prediction accuracy of 85.3% for discriminating the stabilizing and destabilizing protein mutants. For ΔΔG_H2O, we obtained a correlation of 0.78 (standard error 0.96 kcal/mol) with a prediction efficiency of 84.65% in a set of 1603 mutants. We found that the classification of mutants based on secondary structure and ASA remarkably improved the performance, as seen in Figure 2. The correlation improved from 0.52 to 0.85, and the accuracy from 75 to 85% in ΔΔG. Furthermore, we have developed a web server, CUPSAT, for predicting the stability of protein mutants and it is available at http://cupsat.tu-bs.de/ [40].

**Prediction of protein stability from amino acid sequence**

We have developed a method based on interpretable decision tree coupled with adaptive boosting algorithm, and classification and regression tool, respectively for discriminating the stabilizing and destabilizing mutants, and predicting protein stability upon amino acid substitutions just from amino acid sequence. In this method, five variables have been used for discrimination/prediction: (i) Md, mutated (deleted) residue, (ii) Mi, mutant (introduced) residue, (iii) pH, (iv) temperature (°C) at which the stability of the mutated protein was measured explicitly and (v) three neighbouring residues of the central residue. We observed that our method could correctly discriminate the stabilizing and destabilizing protein mutants at an accuracy of 82% in a dataset of 1859 single mutants. Furthermore, a correlation of 0.70 is obtained between the predicted and experimental stabilities. We have set up a web server, iPTREE-STAB, for predicting the stability of proteins and it is available at http://bioinformatics.myweb.hinet.net/iptree.htm [41].

**Conclusions**

A thermodynamic database for proteins and mutants has been set up, which has a large amount of thermodynamic data along with sequence and structure information, experimental methods and conditions, and literature information. The analysis of protein mutant stability revealed that the stability of buried mutations is dominated by hydrophobic interactions, whereas the partially buried and exposed mutations are influenced by hydrophobic, hydrogen bonds and other polar interactions. Furthermore, the classification of mutants based on secondary structures and solvent accessibility could predict the stability of protein mutants with high accuracy. We have proposed different methods for predicting protein stability upon amino acid substitution using structural information, mutated and mutant residues and from amino acid sequence. We have developed web servers for discriminating the stabilizing and destabilizing mutants as well as predicting protein mutant stability, which can be used for any new mutants.

I thank Professor Jenny Littlechild and Professor David Richardson for providing the opportunity to contribute this work to this issue of Biochemical Society Transactions. I also acknowledge Professor A. Sarai, Professor M.N. Ponnuswamy, Dr V. Parthiban, Dr L. Huang and Dr K. Saraboji for stimulating discussions and help at various stages of the work.

**References**
