doi:10.1016/j.jmb.2007.03.069 J. Mol. Biol. (2007) 369, 1318–1332

JMB

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

# The Stability Effects of Protein Mutations Appear to be Universally Distributed

**Nobuhiko Tokuriki[1], Francois Stricher[2], Joost Schymkowitz[3] Luis Serrano[2] and Dan S. Tawfik[1]***

[1]*Department of Biological Chemistry, Weizmann Institute of Science, Rehovot 76100, Israel*

[2]*European Molecular Biology laboratory, Meyerhofstrasse 1 69117 Heidelberg, Germany*

[3]*Vrije Universiteit Brussel Pleinlaan 2, Building E BE-1050 Brussel, Belgium*

**Corresponding author*

How the thermodynamic stability effects of protein mutations ($\Delta\Delta G$) are distributed is a fundamental property related to the architecture, tolerance to mutations (mutational robustness), and evolutionary history of proteins. The stability effects of mutations also dictate the rate and dynamics of protein evolution, with deleterious mutations being the main inhibitory factor. Using the FoldX algorithm that attempts to computationally predict $\Delta\Delta G$ effects of mutations, we deduced the overall distributions of stability effects for all possible mutations in 21 different globular, single domain proteins. We found that these distributions are strikingly similar despite a range of sizes and folds, and largely follow a bi-Gaussian function: The surface residues exhibit a narrow distribution with a mildly destabilizing mean $\Delta\Delta G$ ($\sim$0.6 kcal/mol), whereas the core residues exhibit a wider distribution with a stronger destabilizing mean ($\sim$1.4 kcal/mol). Since smaller proteins have a higher fraction of surface residues, the relative weight of these single distributions correlates with size. We also found that proteins evolved in the laboratory follow an essentially identical distribution, whereas *de novo* designed folds show markedly less destabilizing distributions (i.e. they seem more robust to the effects of mutations). This bi-Gaussian model provides an analytical description of the predicted distributions of mutational stability effects. It comprises a novel tool for analyzing proteins and protein models, for simulating the effect of mutations under evolutionary processes, and a quantitative description of mutational robustness.



*Keywords:* protein stability; mutational robustness; computational biophysics; $\Delta\Delta G$ distributions; protein models

## Introduction

Globular proteins are marginally stable under physiological conditions, with an overall thermodynamic stability ($\Delta G$ folding) in the range of $-5$ to $-15$ kcal/mol.[1] To put these values in context, the energy of single hydrogen bonds is 2–5 kcal/mol. And thus, a single amino acid substitution could dramatically alter the stability of a protein. The comprehensive understanding of the effects of mutations on the stability of proteins is crucial for understanding protein sequence–structure relationships,[2] engineering protein stability,[3,4] simulating and predicting the evolutionary dynamics of proteins,[5–8] validating and refining various protein models and simulations,[9–11] and the *de novo* design of proteins.[12]

Despite the importance of quantitatively understanding the stability effects of mutations, the overall distribution of the $\Delta\Delta G$ effects of mutations is currently unknown. Several comprehensive studies investigated the $\Delta\Delta G$ effects of mutations in proteins such as staphylococcal nuclease[13–17] and barnase.[18,19] These studies show that many, if not most, mutations are destabilizing, and a single point mutation can make a protein completely "collapse". For example, a substitution into a hydrophilic residue in the protein's hydrophobic core is frequently detrimental.[13,20,21] On the other hand, it has also been argued that proteins are tolerant against most mutations,[22–26] and a large number of mutations may be stabilizing.[25,27] Overall, the fraction of mutations that were found to be stabilizing, or destabilizing, varied according to the protein and the nature of these

Abbreviations used: PCA, principal component analysis; ASA, accessible surface area; PDB, Protein Data Bank.

E-mail address of the corresponding author: tawfik@weizmann.ac.il

substitutions, ranging from approximately 8–29% for stabilizing mutations,[25,28] to 4–45% for deleterious mutations.[28] Thus, previous experimental observations suggest that the distribution of $\Delta\Delta G$ effects might be unique for each protein, and no universal rule could explain the differences between proteins, let alone predict such distributions. On the other hand, lattice model proteins showed that, despite different sequences and packing configurations, the $\Delta\Delta G$ distributions for all possible mutations of these model proteins were very similar,[29] at least in their overall shape.[7] However, lattice model distributions can be totally different depending on how the model protein evolved.[25] It is also unclear to what degree the distributions of these model proteins reflect that of real proteins.

In recent years, the energetics of mutant proteins have been studied extensively by both computational and experimental approaches. Several algorithms that predict $\Delta\Delta G$ changes have been developed, and compared with experimental data.[30–39] Amongst these is FoldX, an empirical potential approach that derives an energy function by using a weighted combination of physical energy terms (e.g. van der Waals interactions, hydrogen-bonding, electrostatics, and solvation), statistical energy terms, and structural descriptors, and calibrates these factors to fit experimental $\Delta\Delta G$ values.[30,31] The $\Delta\Delta G$ predictions by FoldX were validated using a large set of mutations in a range of different real proteins. The utility of FoldX in designing thermostable proteins,[40,41] and in predicting the effects of mutations on binding energies,[42] and fitness changes of proteins,[7,8] has also been demonstrated.

Here, we applied FoldX to predict the $\Delta\Delta G$ values for all possible mutations in 21 different proteins. We obtained the computational distributions of $\Delta\Delta G$ effects of all mutations in these proteins, compared them to experimental values available for a partial set of mutations in a number of these proteins, and extrapolated several universal rules that may account for, and possibly predict, such distributions. Although the FoldX values are a prediction and obviously have limited accuracy, they enabled us to examine $\Delta\Delta G$ distributions in a protein-based physical model. Thus, whilst the values for individual mutations can considerably deviate from the experimental values, the trends that we observed are likely to be relevant to real proteins.[43]

## Results

### Validation of FoldX computed distributions

The thermodynamic stability changes of mutations were computed using the force-field FoldX (version 2.52). We followed a four-step procedure as described.[44] First, protein structures (previously determined by X-ray crystallography) were optimized using the repair function of FoldX. Second, structures corresponding to each of the single point mutants (self-mutated structures) were generated by the repair position scan function of FoldX. Third, the energies for these structures were calculated using the energy calculation function of FoldX. Fourth, the energy values of the mutant structure were compared with those of the wild-type structures.

FoldX has been optimized for speed and applicability, and several changes have been made in the energy calculations since the original version was reported. We therefore revalidated the $\Delta\Delta G$ values computed by FoldX by comparing them to data from 1285 experimentally measured mutants of ten different proteins available from the ProTherm database† (Supplementary Data Figure 1). Although the entire range of mutations is not available for a single protein, the experimental data are very helpful in validating the FoldX predictions. In addition, in the early version of FoldX, only certain tendencies of mutations, such as the removal of groups from side-chains, were considered. Here, all types of mutations were tested, including mutations from a small into a larger side-chain, both on the surface and within the proteins' core (F.S. and L.S., unpublished results).

The correlation of the FoldX and experimental values was previously based on linear regression.[30] Here we examined the correlation of the calculated and experimental values by linear regression, as well as principal component analysis (PCA), which better addresses complex and large datasets. The $\Delta\Delta G$ values calculated by FoldX for the ProTherm set of experimental mutations were normalized using either the linear, or the PCA, function, and presented in histograms by classifying 25 bins, each 1.0 kcal/mol wide (Supplementary Data Figure 1). The computed FoldX values (with no normalization) gave a distribution that is quite similar to that of the experimental values, and the normalization by the PCA correlation led to essentially identical distributions (Supplementary Data Figure 2; Figure 1). In contrast, the distribution of values normalized by the linear equation significantly deviated from the distribution of the experimental values. Subsequently, all FoldX values were corrected using the PCA equation ($\Delta\Delta G^{FoldX} = -0.078 + 1.14\Delta\Delta G^{Experimental}$; Supplementary Data Figure 1), although in effect, under the subtle correction of the PCA equation, the vast majority of values (94%) remained within error range of the directly computed values with no normalization (±0.5 kcal/mol).

The systematic comparison of the computed *versus* the experimental values along a large set of mutations of different types generally revealed a consistent correlation, although certain tendencies, or biases, were observed. Most notably, the stabilizing effects of mutations into hydrophilic residues (Arg and Asp, primarily) tend to be overestimated. However, it was found that the vast majority of mutations were distributed evenly around the linear equation obtained by PCA (F.S. and L.S., unpublished results).

† http://gibk26.bse.kyutech.ac.jp/jouhou/protherm/protherm.html