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 $\Delta\Delta G$ (~0.6 kcal/mol), whereas the core residues exhibit a wider distribution with a stronger destabilizing mean (~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

## The ΔΔ*G* distributions of natural proteins

We have initially explored 16 natural, single domain, monomeric proteins (with the exception of barnase, which is a trimer) with different folds and sizes (50–330 chain length), for which crystal structures are available with relatively high resolution (Table 1). These were mostly enzymes, including enzymes that are heavily represented in the experimental dataset that was used to calibrate the FoldX values (see previous section).[7,8,23,45–49] The ΔΔ*G* values for all possible mutations in each of these proteins were calculated by FoldX, and presented as histograms (Figure 1). All mutations attainable by single nucleotide substitutions were also plotted. This is because in nature, the majority of codon changes are initially limited to single nucleotide substitutions, thus limiting the diversity of amino acid exchanges attainable through immediate mutational changes.

Despite having different folds and chain lengths, all 16 proteins exhibited a similar distribution. Interestingly, the 1285 experimental mutations dataset exhibits a similar distribution. However, this observation must be considered in view of the fact that these mutations belong to ten different proteins, and the type of mutations is often biased. The most frequent mutations are mildly deleterious (~1 kcal/mol), both in all, and single nucleotide, mutations. The distributions are all asymmetric with a sharp slope leading to −2 kcal/mol, and a shoulder towards 7 kcal/mol. Such asymmetric distributions of ΔΔ*G* were also observed in the studies of the lattice model proteins.[7,25,29] On average, the distributions of single nucleotide substitution have less (12% *versus* 15%) highly destabilizing mutations ($\Delta\Delta G > 3$ kcal/mol), and more (48% *versus* 44%) neutral mutations ($-1 < \Delta\Delta G < 1$ kcal/mol). This is consistent with the known fact that codons related by single nucleotide exchanges tend to code a similar type of amino acid.[50] Consequently, the average ΔΔ*G* of all possible mutations is slightly more destabilizing than for single nucleotide mutations (by 0.12 kcal/mol, on average; Table 1).

## The bi-Gaussian model of ΔΔ*G* distributions

We subsequently attempted to describe these ΔΔ*G* distributions by a simple function. Tiana and coworkers described the ΔΔ*G* distributions of lattice model proteins with a bi-Gaussian function, and found that this function was almost identical for different sequences and conformations.[29] We have attempted to fit the FoldX distributions to the same function (equation (1)):

$$F_{\text{bi-Gaussian}}(x) = 100\left\{\frac{P_1}{\sqrt{2\pi \times \sigma_1{}^2}}\exp\left[-(x-\mu_1)^2/2\sigma_1{}^2\right] + \frac{1-P_1}{\sqrt{2\pi \times \sigma_2{}^2}}\exp\left[-(x-\mu_2)^2/2\sigma_2{}^2\right]\right\} \tag{1}$$

where $F(x)$ is a percentage-based probability distribution function, $P_1$ ($0<P_1<1$) is the fraction of one Gaussian function (and ($1–P_1$) is therefore the fraction of the second Gaussian function), $\mu_1$ and $\mu_2$, are the mean values of each Gaussian function, and $\sigma_1$ and $\sigma_2$, the corresponding standard deviations.

The bi-Gaussian model provided an excellent description of the FoldX distributions, and in particular those of single nucleotide mutations ($R \geq 0.99$) (Figure 2(a) and Supplementary Data Figure 3a). Interestingly, the individual Gaussian distributions derived from these fits are quite similar for all proteins tested: One Gaussian has a mildly deleterious mean value ($\mu_1 = 0.56 \pm 0.12$) and a sharp distribution ($\sigma_1 = 0.90 \pm 0.16$); the other exhibits a stronger destabilizing mean ($\mu_2 = 1.96 \pm 0.53$) and a broader distribution ($\sigma_2 = 1.93 \pm 0.29$) (Table 2). These similarities suggest that ΔΔ*G* distribution of proteins can be described in more general terms, such that the bi-Gaussian function uses these average μ and σ values. The individual Gaussian values were thus fixed to the average values ($\mu_1 = 0.56$, $\sigma_1 = 0.90$, $\mu_2 = 1.96$, $\sigma_2 = 1.93$), and only $P_1$ was acquired through fitting to equation (2):

$$F(x) = 100\left\{\frac{P_1}{\sqrt{2\pi \times 0.90^2}}\exp\left[-(x-0.56)^2/2\times 0.90^2\right] + \frac{1-P_1}{\sqrt{2\pi \times 1.93^2}}\exp\left[-(x-1.96)^2/2\times 1.93^2\right]\right\} \tag{2}$$

where $P_1$ is in the range of 0 to 1.

The fits obtained for equation (2) were quite good ($R \geq 0.99$, Figure 2(b) and Supplementary Data Figure 3b). In the same way, the mean values, and standard deviations, of the bi-Gaussian distributions of all possible mutations have been acquired (Supplementary Figure 4a), yielding the following average values: $\mu_1 = 0.54 \pm 0.15$, $\sigma_1 = 0.98 \pm 0.12$, $\mu_2 = 2.05 \pm 0.36$, $\sigma_2 = 1.91 \pm 0.22$ (Table 2). These distributions were then fitted to the bi-Gaussian model with these average values to give equation (2′) (Supplementary Figure 4b):

$$F(x) = 100\left\{\frac{P_1}{\sqrt{2\pi \times 0.98^2}}\exp\left[-(x-0.54)^2/2\times 0.98^2\right] + \frac{1-P_1}{\sqrt{2\pi \times 1.91^2}}\exp\left[-(x-2.05)^2/2\times 1.91^2\right]\right\} \tag{2′}$$

The ΔΔ*G* distribution of the experimental dataset fit quite well to both equations (1) and (2′) (Figure 2), and show μ and σ values that are similar to the average values obtained by analyzing 16 different proteins with FoldX (Table 2, lower panel).

The bi-Gaussian descriptions indicate that the distribution of ΔΔ*G* effects, as computed by FoldX and supported by the experimental data, follows a

**Table 1.** Summary features of the studied proteins

| Protein | | | | | | Average ΔΔ*G* (kacl/mol)[c] | |
|---|---|---|---|---|---|---|---|
| Common name | Abbreviation | Chain length (no. amino acids) | SCOP classification[a] | PDB code | Average ASA[b] | All possible mutations | Single nucleotide mutations |
| Recombinant serum paraoxonase 1 | PON | 332 | 6-bladed beta propeller | 1V04 | 0.277 | 1.54 | 1.39 |
| Lipase | Lipase | 285 | Alpha/beta hydrolase | 1EX9 | 0.274 | 1.26 | 1.13 |
| β-Lactamase | TEM1 | 263 | beta-lactamase/transpeptidase-like | 1BTL | 0.282 | 1.32 | 1.14 |
| Human carbonic anhydorase II | CAII | 259 | Carbonic anhydrase | 1LUG | 0.298 | 1.60 | 1.44 |
| Dihydrofolate reductase | DHFR | 159 | Dihydrofolate reductases | 1RX2 | 0.337 | 1.31 | 1.04 |
| Robinuclease H | RNase H | 155 | Ribonuclease H-like | 2RN2 | 0.336 | 1.16 | 1.17 |
| Myoglobin | Myoglobin | 151 | Globin-like | 1A6K | 0.334 | 1.10 | 0.95 |
| Staphilococcus nuclease | SNase | 136 | OB-fold | 1STN | 0.332 | 1.36 | 1.21 |
| Human lysozome | Human lysozome | 130 | Lysozyme-like | 1REX | 0.331 | 1.60 | 1.54 |
| Hen lysozome | Hen lysozome | 129 | Lysozyme-like | 1DPX | 0.339 | 1.74 | 1.67 |
| Ribonuclease A | RNase A | 124 | RNase A-like | 1FS3 | 0.368 | 1.33 | 1.32 |
| Barnase | Barnase | 108 | Microbial ribonucleases | 1A2P | 0.356 | 1.52 | 1.41 |
| Acylphosphatase | AcP | 98 | Ferredoxin-like | 2ACY | 0.357 | 1.40 | 1.17 |
| Ubiquitin | Ubiquitin | 76 | beta-Grasp (ubiquitin-like) | 1UBQ | 0.411 | 1.07 | 0.83 |
| Protein G | Protein G | 61 | beta-Grasp (ubiquitin-like) | 2IGD | 0.454 | 0.93 | 0.92 |
| Cro repressor | Cro repressor | 59 | Lambda repressor-like DNA-binding domains | 1ORC | 0.442 | 1.09 | 0.98 |
| Average | | | | | | 1.33 | 1.21 |
| *Novel proteins* | | | | | | | |
| Ankyrin repeat protein | Ankyrin repeat protein | 156 | Artificial ankyrin repeat proteins | 1MJ0 | 0.310 | 1.37 | |
| Nevel fold-computationally designed | TOP7 | 92 | New fold designs | 1QYS | 0.364 | 0.80 | |
| Combnation protein 1B11 | 1B11 | 86 | *In vitro* evolution products | 2NH8 | 0.422 | 1.36 | |
| A novel fold from *in vitro* evolution | ADBP | 67 | *In vitro* evolution products | 1UW1 | 0.460 | 1.31 | |
| Redesigned protein G | Redesigned protein G | 57 | beta-Grasp (ubiquitin-like) | 1MHX | 0.455 | 0.74 | |

[a] SCOP definition was derived from the Structural Classification of Proteins database [http://www.scop.mrc-lmb.cam.ac.uk/scop/].

[b] Average ASA values correspond to the average of the surface accessibility values (ASA) of all residues in a given protein.

[c] Average ΔΔ*G* values correspond to the average of ΔΔ*G* values of the entire set of the protein's single nucleotide mutations, or all possible mutations.

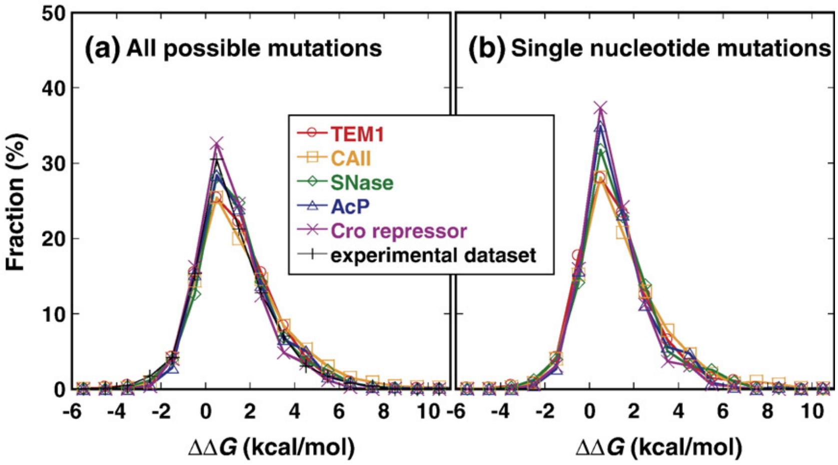

**Figure 1.** The $\Delta\Delta G$ distributions of several natural proteins (for details see Table 1). The $\Delta\Delta G$ values of all possible mutations, at each amino acid position along the chosen protein, were computed by FoldX. The data are presented in histograms, using 1 kcal/mol bins, from $-10$ kcal/mol to 15 kcal/mol (the few mutations with $\Delta\Delta G > 14$ kcal/mol were classified into the 14–15 kcal/mol bin, and the very few mutations with $\Delta\Delta G < -9$ kcal/mol into the $(-10)$–$(-9)$ bin). (a) The distribution of all possible mutations. The FoldX computed distribution of $\Delta\Delta G$ values of all 19 possible amino acid substitutions per each position, and the distribution of the experimental $\Delta\Delta G$ values for the dataset of 1285 mutations. (The datasets of mutations presented here are not identical; a comparison of the distributions for same set of mutation gives similar results and is available as Supplementary Data Figure 2). (b) The distribution of single nucleotide mutations. The $\Delta\Delta G$ values of all mutations afforded by single nucleotide substitutions of the gene encoding the presented proteins were selected from the pool of all possible mutations, and distributed into bins as above.

universal rule, and is largely independent of sequence composition and fold. By and large, all proteins tested follow the bi-Gaussian function presented in equations (2), or (2′). The most systematically variable parameter seems to be $P_1$, i.e. the relative fraction of each distribution (Table 2).

## Correlating ΔΔG with accessible surface area

Why can the $\Delta\Delta G$ distributions of proteins be expressed as the superposition of two Gaussian distributions? Tiana and co-workers showed that the two Gaussian distributions of lattice model proteins stemmed from “hot” and “cold” sites in relation to protein folding.[29] We surmised that proteins are composed of a hydrophobic core, and a hydrophilic surface (an “oil droplet in water”).[51] The core plays a key role in protein folding and stability, and core mutations are considered more deleterious than surface mutations.[24] Thus, the two Gaussian distributions may relate to core and surface residues. To separate the core from the surface, we applied accessible surface area values (ASA)[52] that, based on the 3D structure, indicate to what extent an amino acid residue is exposed to the solvent. Indeed, there is a clear correlation between the ASA of residues, and the $\Delta\Delta G$ effects of mutations in these residues (Figure 3). Most of the highly destabilizing mutations ($\Delta\Delta G > 5$) are located in the core ($\text{ASA} < 0.25$). Amongst surface residues ($\text{ASA} \geq 0.25$), there are hardly any highly destabilizing mutations, and the average $\Delta\Delta G$ is around 0.5 kcal/mol. Notably, the $\Delta\Delta G$ values of the experimental dataset showed the same tendency. It should be noted, however, that the ASA analysis revealed that certain types of mutations show poor predictions. In particular, mutations into hydrophilic residues such as Arg and Asp are predicted to be highly stabilizing, but this seems to be an overestimation of FoldX, since these mutations have no parallels in the experimental dataset (see also Supplementary Figure 1). However, the contribution of these few mutations to the overall distribution is negligible.

The correlation between $\Delta\Delta G$ and solvent accessibility suggested that the two individual Gaussians may correspond to these two parts of the protein, namely core and surface. Thus, protein residues were classified into two categories according to the ASA values. We applied different ASA cut-offs (0.1, 0.5), and for each cut-off attempted to describe the distributions of the resulting groups of surface and core residues, each by a separate Gaussian function:

$$F(x) = \frac{100}{\sqrt{2\pi\sigma^2}} \exp\left[-(x-\mu)^2/2\sigma^2\right] \qquad (3)$$

where μ is the mean, and σ is the standard deviation.

Around a cut-off of 0.25, the mono-Gaussian distributions of the proteins tested showed the best fit for both the surface and core (Figure 4 and Supplementary Data Table 1). For surface residues ($\text{ASA} \geq 0.25$), the distributions exhibited nearly neutral means ($\mu = 0.59 \pm 0.11$, $\sigma = 1.09 \pm 0.10$), and the fit was good ($R > 0.99$). For core residues ($\text{ASA} < 0.25$), the distributions had stronger destabilizing means ($\mu = 1.34 \pm 0.21$, $\sigma = 1.74 \pm 0.20$), and the fit was generally poorer ($R = 0.95$–$0.99$). The 1285 mutants of the experimental dataset and the $\Delta\Delta G$ values computed by FoldX showed the same tendency (Figure 4). The individual distributions for surface and core residues are therefore similar to those obtained with the bi-Gaussian model (Table 2). Moreover, for the entire set of proteins, the fraction of surface residues attained by the fit to the bi-Gaussian model ($P_1$ in equation (2)) correlates very well with the fraction of surface residues that possess ASA values that are $\geq 0.25$ (Figure 5(a)).

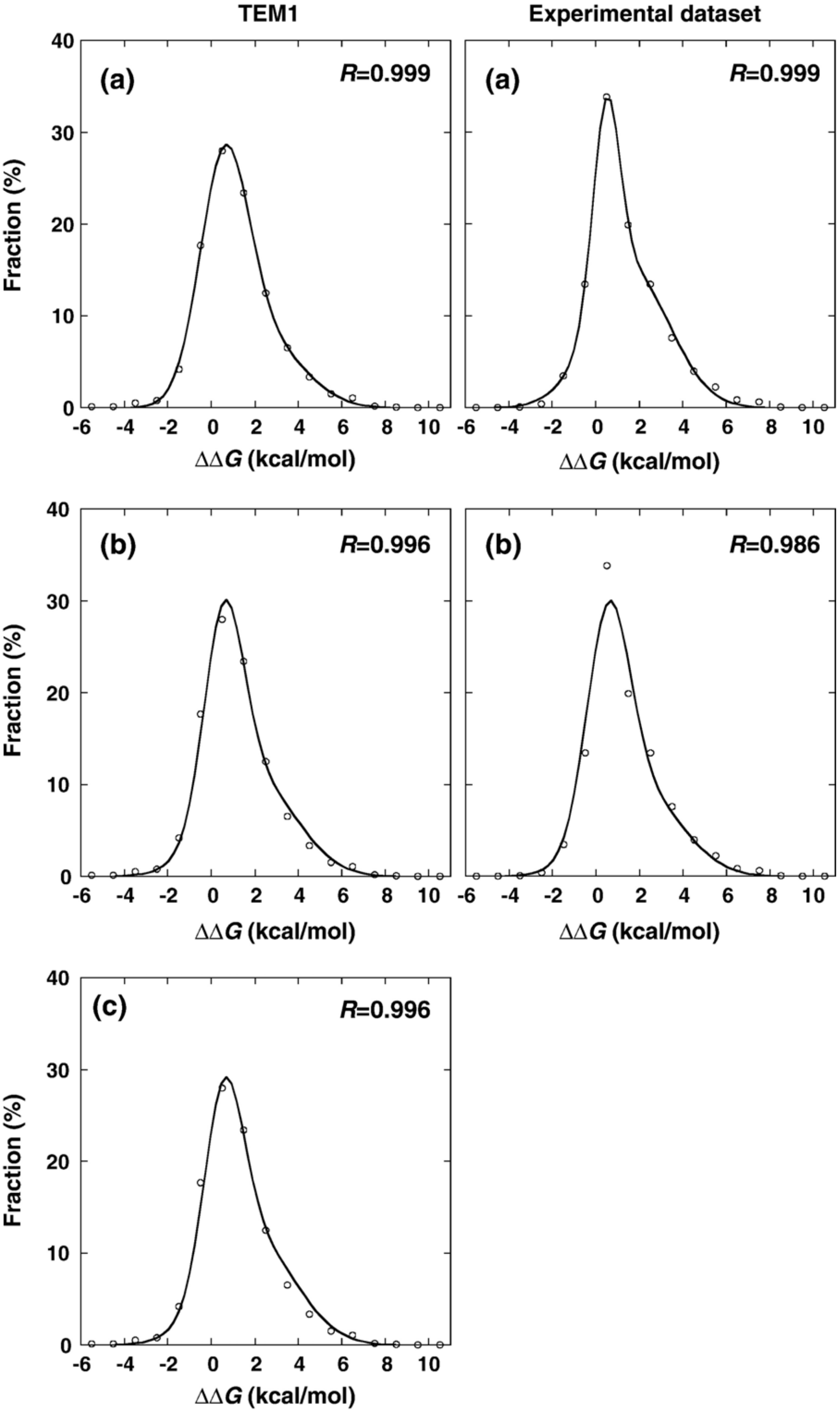

**Figure 2.** The bi-Gaussian model of $\Delta\Delta G$ distributions. (a) The FoldX computed distribution of all single nucleotide mutations of a representative protein (TEM-1), and the distribution of the experimental dataset of 1285 mutations, fitted to equation (1). The resulting parameters are provided in Table 1. (The fits for all other proteins are provided as Supplementary Data Figure 3a). (b) The same distributions fitted to equation (2). (The fits for all other proteins are provided as Supplementary Data Figure 3b). (c) The TEM-1 distribution fitted to the universal model: equation (2) was applied (with the same mean values for the individual Gaussians as in (b)) while deriving $P_1$ from TEM-1 chain length (equation (4)). (The fits for all other proteins are provided as Supplementary Data Figure 3c; the experimental dataset is comprised of ten different proteins each with a different chain length, and is therefore inadequate for this model).

**Table 2.** Summary of mean (μ), distribution (σ) and partition ($P_1$) values of the studies proteins

| Protein | | Single point mutations | | | | | | All possible mutations | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Common name | Abbreviation | $\mu_1$ | $\sigma_1$ | $\mu_2$ | $\sigma_2$ | $P_1$[a] | $R$[a] | $\mu_1$ | $\sigma_1$ | $\mu_2$ | $\sigma_2$ | $P_1$[a] | $R$[a] |
| Recombinant serum paraoxonase 1 | PON | 0.51 | 0.91 | 1.93 | 1.82 | 0.42 | 1.000 | 0.64 | 0.91 | 1.93 | 2.01 | 0.34 | 1.000 |
| Lipase | Lipase | 0.57 | 1.18 | 2.27 | 2.05 | 0.67 | 0.999 | 0.47 | 1.21 | 2.25 | 2.07 | 0.60 | 1.000 |
| β-Lactamase | TEM1 | 0.58 | 1.11 | 2.36 | 1.84 | 0.67 | 0.999 | 0.44 | 0.96 | 1.69 | 1.77 | 0.32 | 0.999 |
| Human carbonic anhydorase II | CAII | 0.50 | 0.85 | 1.80 | 1.99 | 0.39 | 0.999 | 0.48 | 0.88 | 1.92 | 2.03 | 0.34 | 0.998 |
| Dihydrofolate reductase | DHFR | 0.54 | 0.73 | 1.53 | 1.78 | 0.39 | 0.999 | 0.43 | 0.93 | 1.98 | 1.78 | 0.43 | 1.000 |
| Robinuclease H | RNase H | 0.49 | 0.98 | 2.12 | 1.98 | 0.69 | 1.000 | 0.48 | 0.95 | 1.95 | 1.96 | 0.56 | 1.000 |
| Myoglobin | Myoglobin | 0.31 | 0.84 | 1.53 | 1.57 | 0.48 | 1.000 | 0.27 | 1.02 | 1.95 | 1.46 | 0.51 | 1.000 |
| Staphilococcus nuclease | SNase | 0.58 | 0.68 | 1.30 | 1.71 | 0.30 | 0.997 | 0.80 | 0.92 | 1.74 | 2.04 | 0.46 | 0.999 |
| Human lysozome | Human lysozome | 0.76 | 1.12 | 3.02 | 2.29 | 0.70 | 0.999 | 0.71 | 1.19 | 2.44 | 2.06 | 0.54 | 0.998 |
| Hen lysozome | Hen lysozome | 0.81 | 1.16 | 3.00 | 2.43 | 0.68 | 0.998 | 0.77 | 1.16 | 2.91 | 2.09 | 0.59 | 0.999 |
| Ribonuclease A | RNase A | 0.59 | 0.83 | 1.76 | 2.56 | 0.55 | 0.998 | 0.59 | 1.01 | 2.12 | 2.39 | 0.55 | 1.000 |
| Barnase | Barnase | 0.59 | 0.84 | 2.08 | 1.93 | 0.51 | 0.999 | 0.49 | 0.82 | 1.62 | 1.73 | 0.43 | 1.000 |
| Acylphosphatase | AcP | 0.56 | 0.80 | 1.93 | 1.83 | 0.56 | 0.999 | 0.66 | 1.05 | 2.63 | 1.77 | 0.65 | 0.999 |
| Ubiquitin | Ubiquitin | 0.42 | 0.83 | 1.33 | 1.64 | 0.52 | 1.000 | 0.28 | 0.90 | 1.82 | 1.69 | 0.47 | 1.000 |
| Protein G | Protein G | 0.59 | 0.84 | 2.08 | 1.93 | 0.51 | 0.999 | 0.56 | 0.90 | 2.09 | 1.85 | 0.43 | 1.000 |
| Cro repressor | Cro repressor | 0.55 | 0.74 | 1.33 | 1.58 | 0.48 | 0.999 | 0.61 | 0.92 | 1.68 | 1.78 | 0.57 | 0.999 |
| Average | | 0.56 | 0.90 | 1.96 | 1.93 | 0.53 | | 0.54 | 0.98 | 2.05 | 1.91 | 0.49 | |
| Standard deviation | | 0.12 | 0.16 | 0.53 | 0.29 | 0.12 | | 0.15 | 0.12 | 0.36 | 0.22 | 0.10 | |
| *Novel proteins* | | | | | | | | | | | | | |
| Ankyrin repeat protein | Ankyrin repeat protein | | | | | | | 0.93 | 1.18 | 3.86 | 1.36 | 0.85 | 1.000 |
| Nevel fold-computationally designed | TOP7 | | | | | | | 0.17 | 0.95 | 1.41 | 1.85 | 0.49 | 1.000 |
| Combnation protein 1B11 | 1B11 | | | | | | | 0.61 | 1.07 | 2.51 | 1.72 | 0.62 | 1.000 |
| A novel fold from *in vitro* evolution | ADBP | | | | | | | 0.00 | 0.33 | 1.04 | 1.33 | 0.26 | 0.999 |
| Redesigned protein G | Redesigned protein G | | | | | | | 0.16 | 0.90 | 1.38 | 2.14 | 0.57 | 0.999 |
| *Experimental dataset*[b] | | | | | | | | | | | | | |
| Actual values | | | | | | | | 0.48 | 0.61 | 1.61 | 1.77 | 0.40 | 1.000 |
| FoldX prediction | | | | | | | | 0.58 | 0.71 | 1.64 | 1.73 | 0.42 | 0.998 |

These parameters were derived from fitting the $\Delta\Delta G$ distributions to equation (1).

[a] The FoldX computed distributions were fitted to equation (1), and the resulting parameters are noted. Also noted is the correlation coefficient ($R$). For examples, of such fits, see Figure 2(a), all other fits are provided in Supplementary Data Figures 3(a) and 4(a).

[b] The parameters related to the fit of the distribution of $\Delta\Delta G$ values for the 1285 mutations dataset.

## A universal function describing $\Delta\Delta G$ distributions

Thus, by a reasonable approximation, the two individual distributions obtained by the bi-Gaussian model represent the distribution of $\Delta\Delta G$ values for the core, and surface, residues. It also seems that the fraction of surface residues ($P_1$) correlates with protein size. Indeed, the fit of $P_1$ values to the log of chain length, or number of amino acids ($L$), gave rise to equation (4) (for single nucleotide substitutions) and (4′) (for all possible mutations):

$$P_1 = 1.27 - 0.33\log L \qquad (4)$$

$$P_1 = 1.13 - 0.30\log L \qquad (4')$$

where $P_1$ (the fraction of the first Gaussian) takes values between 0 and 1; and $L$ for the proteins described here is 50–330 amino acid residues.

Given that the average mean values (μ) and distribution widths (σ) for the core, and surface, residues can also be applied (Figure 2(b)), the $\Delta\Delta G$ distribution of a protein could be largely described by combining equation (2) (for single nucleotide mutations), or (2′) (for all possible mutations), with equation (4), or (4′), respectively. As seen in Figure 2(c), and Supplementary Data Figures 3c and 4c, the distributions of all 16 natural proteins examined here are described by this model with reasonable accuracy ($R \geq 0.98$), with the only required input being the protein's chain length ($L$).

## The $\Delta\Delta G$ distribution of novel proteins

Natural proteins possess a long history of evolution. Hence, the universal distribution presented above could be the consequence of random drift and natural selection, or it may reflect an inherent property shared by all globular proteins. Over the past decade, *in vitro* evolution, and rational and computational design were applied towards the generation of novel proteins. Would these novel proteins exhibit the same $\Delta\Delta G$ distributions? We have investigated five different novel proteins generated by *in vitro* evolution,[53–55] rational design of a new scaffold,[56] and computational design[57–59] (Table 1). The average ASA values of all these proteins were well correlated with their size, as observed for natural proteins (Figure 5(c)). This indicated that novel and natural proteins are likely to have similar packing of protein core and surface. Three proteins (an engineered ankyrin repeat protein,[56] combinational protein 1B11 obtained by combinatorial shuffling of polypeptide segments grafted from existing proteins,[55] and ANBP, a novel fold obtained by selection from a library of

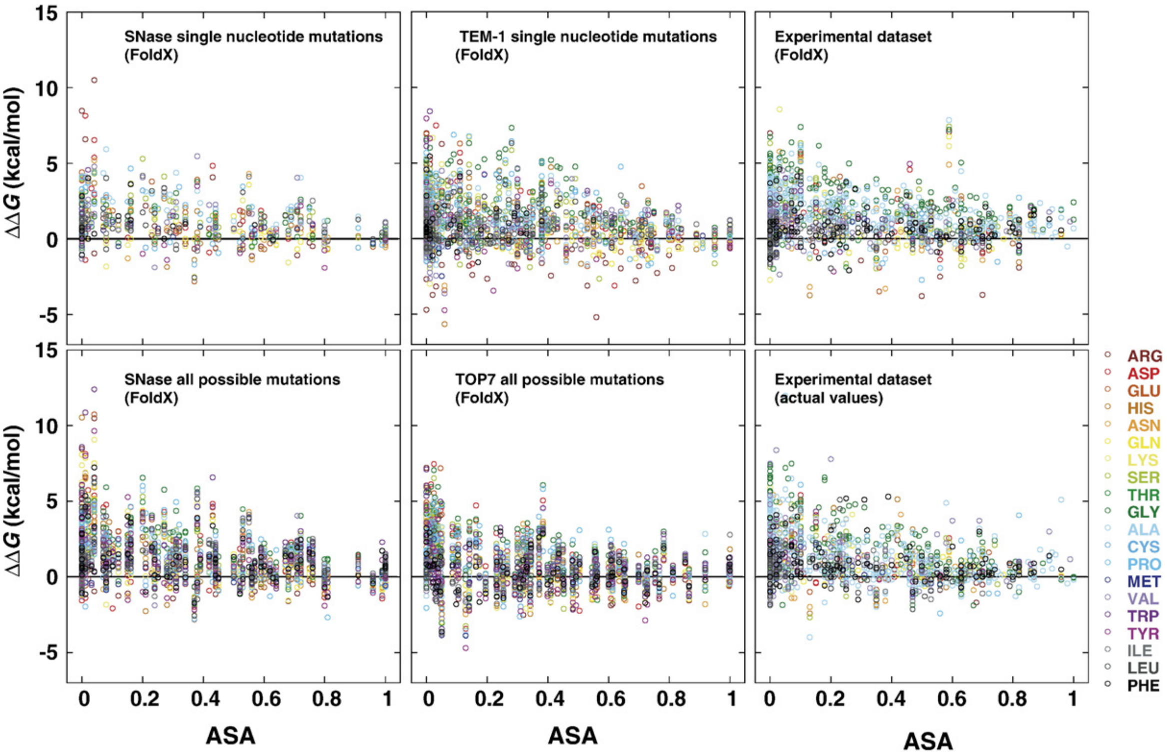

**Figure 3.** $\Delta\Delta G$ values as a function of solvent accessibility. Presented, for each amino acid along the protein's chain, is the accessible solvent area of that amino acid (ASA), and the $\Delta\Delta G$ values for all possible mutations at this position. The color codes for the mutants' amino acids are indicated (i.e. the various amino acids that the noted position was mutated into). Presented are four representative proteins analyzed by FoldX, and the ProTherm experimental dataset with both the experimentally measured values and the FoldX predictions for the same mutations.

completely random sequences[53]) also showed similar $\Delta\Delta G$ distributions to those of natural proteins. However, two proteins obtained by computational design, TOP7[57] and a redesigned protein G,[58,59] showed a different distribution (Figure 6 and Table 2). Unfortunately, the gene sequences of these proteins are not available, and hence the distributions of "single nucleotide mutations", that show much better fit to the universal model, could not be computed. Nevertheless, in comparison to natural proteins, these computationally designed proteins have a higher fraction of stabilizing mutations, and a much lower fraction of destabilizing mutations (Figure 6), resulting in a mean $\Delta\Delta G$ value that is more stabilizing than that of natural proteins of equivalent size (Table 1 and Figure 5(d)).

This comparison between novel man-made proteins and natural ones, although based on a rather small number of novel proteins for which a 3D structure is available, suggests that the bi-Gaussian distributions of $\Delta\Delta G$ values according to core and surface are an inherent property of globular proteins. However, the mean values for each distribution might be related to the protein's origin. Interestingly, the impact of both natural and artificial selection seems to be similar. A novel protein selected in the laboratory from a library of completely random sequences[53,54] exhibits a distribution similar to proteins that have been under natural selection for many millions of years (Figure 6, ANBP and 1B11). In contrast, computationally designed proteins show a much more "robust" distribution, by which, the deleterious effects of mutations are significantly minimized (Figure 6, TOP7, and redesigned protein G).

## Discussion

### Predicting ΔΔ*G* distributions with FoldX

Computational methods have been much improved in the last several years, but these methods are yet incapable of predicting $\Delta\Delta G$ values in perfect accuracy. It is especially difficult to predict $\Delta\Delta G$ values for mutations that cause conformational changes with force fields such as FoldX that assume a fixed backbone. There are also certain tendencies, or biases, related to a particular type of mutation. These biases, however, are relatively minor, and seem largely negligible for the analysis of large datasets such as the overall distributions of $\Delta\Delta G$ values. Here, we also classified the individual $\Delta\Delta G$ values into 1 kcal/mol wide bins, that are largely within the expected error range of FoldX. To further validate the FoldX predictions, we have compared them with a large dataset of 1285 mutations with experimentally available $\Delta\Delta G$ values. Although the experimental dataset relates to ten different proteins, and the choice of mutations is often biased, their overall distributions are compatible with those

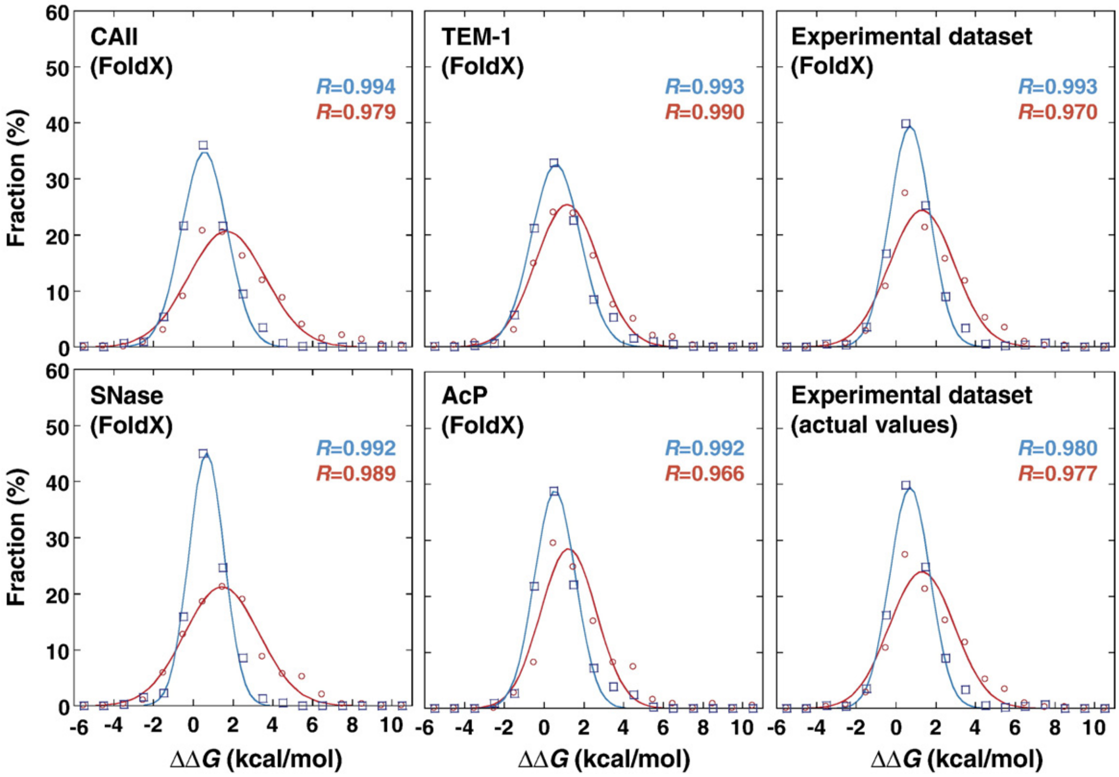

**Figure 4.** The individual $\Delta\Delta G$ distributions of core and surface residues. The residues of each protein were divided according to their ASA values: core (ASA < 0.25; in red) and surface (ASA ≥ 0.25; in blue). The $\Delta\Delta G$ values for single nucleotide mutations are presented in histograms, using 1 kcal/mol bins as above (Figure 1). The distributions were fitted to a single Gaussian function (equation (3)). Presented are four representative proteins analyzed by FoldX, and the ProTherm experimental dataset with both the experimentally measured values and the FoldX predictions for the same mutations. The fits of all other proteins, the mean values (μ), standard deviations (σ), and correlations (*R* values) of these fits, are summarized in Supplementary Data Table 1.

obtained with FoldX. The experimental $\Delta\Delta G$ distribution was well described by a bi-Gaussian with mean values (μ) and widths (σ) that are similar to the average values of 16 proteins analyzed by FoldX (Figures 1 and 2; Table 2). The separated $\Delta\Delta G$ distributions for surface and core residues also followed mono-Gaussians with values similar to those obtained with FoldX (Figures 3 and 4; Supplementary Data Table 1). Furthermore, $\Delta\Delta G$ computations of ~1000 different mutations that accumulated under random mutational drift in one of the proteins analyzed here (TEM-1) indicated a remarkable correlation between the $\Delta\Delta G$ values of the mutation and its tolerance under a given selection pressure.[8] Despite all these evidences in support of the accuracy of FoldX predictions, inaccuracies, and biases, that can affect the reported $\Delta\Delta G$ distributions, and the values instated in our model, are obviously inevitable. However, as previously noted,[43] whilst the one-to-one comparisons of computed and experimental $\Delta\Delta G$ values indicate considerable deviations, the computational predictions seem to capture the overall trends in a strikingly reliable manner.

Another reassuring factor is that, on the whole, our findings are consistent with known general properties of proteins. The $\Delta\Delta G$ distributions of all possible mutations are more destabilizing than that of single nucleotide mutations;[50] and, on average, mutations in core residues are much more destabilizing than mutations on the surface.[20,24,60]

## A universal distribution of ΔΔ*G*

The FoldX-based analysis revealed that evolved proteins, both in nature and in the laboratory, show very similar distributions of $\Delta\Delta G$ effects, independent of their sequence and fold. As indicated above, this distribution could be expressed by a bi-Gaussian function, the only input parameter of which is chain length (equations (3) and (4)). The universal distribution of $\Delta\Delta G$ values implies that the folding and stability of globular proteins is governed by simple rules. The mutations on the surface are almost never highly destabilizing, and generally deviate around neutrality, whilst mutations of core residues have a broad distribution and a larger destabilizing mean.

The analysis of novel proteins revealed several interesting aspects of the $\Delta\Delta G$ distributions and their dependency on the origin of these proteins. First, the distribution of novel proteins selected in the laboratory by several rounds of mutation and selection appears to be identical to the distributions of proteins that had been under natural selection for many millions of years (Figure 6). Second, as evident

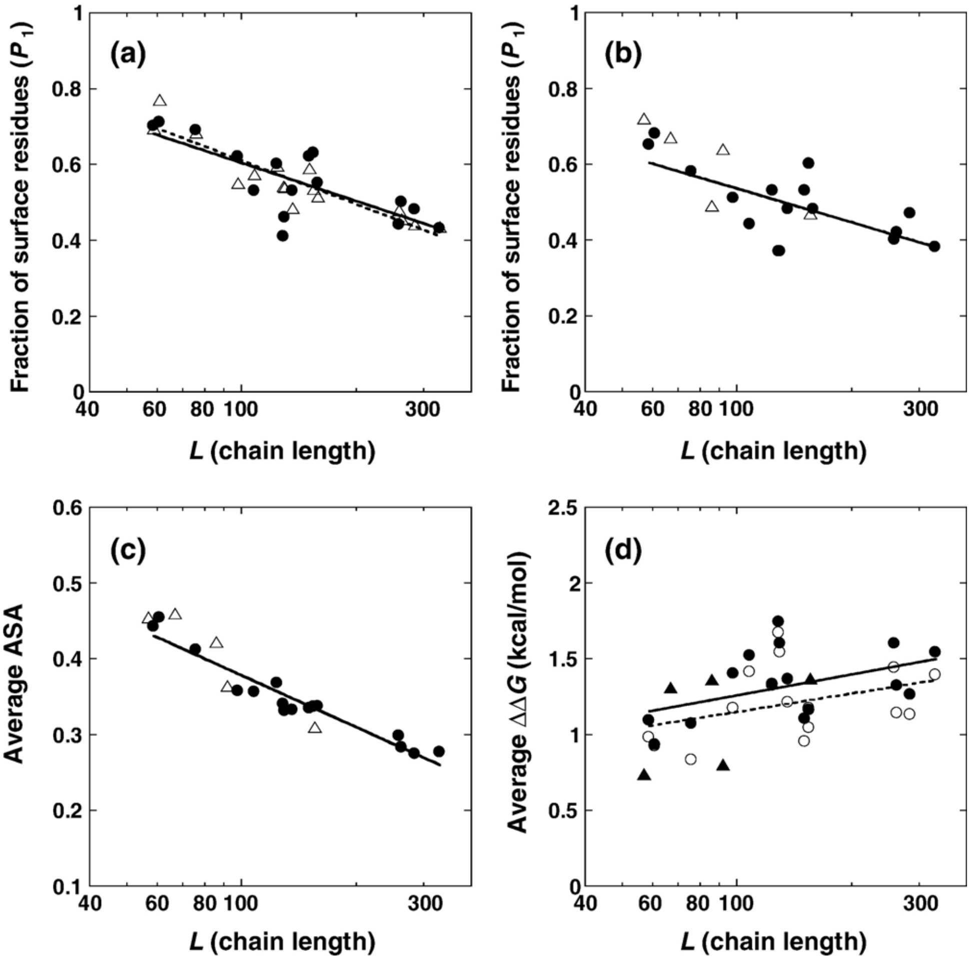

**Figure 5.** The correlation between the chain length of proteins, and their properties. (a) The correlation between chain length (number of amino acids) and $P_1$ (fraction of surface residues) for single nucleotide mutations. One set of $P_1$ values corresponds to the fraction of residues possessing an ASA value that is ≥0.25 (△; broken line). The other set (●; continuous line) was derived from the fit of ΔΔ*G* distributions to a bi-Gaussian function using average μ and σ values (equation (2)). The fit of this continuous line yielded equation (4) ($P_1 = 1.27–0.33\log L$). (b) The correlation between $P_1$ for all possible mutations, and chain length. Filled circles (●) (continuous line) are for $P_1$ of natural proteins, which was derived from the fit of ΔΔ*G* distributions to a bi-Gaussian function using average μ and σ values (equation (2′)). Open triangles (△) are for novel proteins. The continuous line represents a fit to equation (4′) $P_1 = 1.13–0.30\log L$. (c) The correlation between proteins' chain length and their average surface accessibility value. Filled circles (●) are for natural proteins, open triangles (△) for novel proteins. (d) The correlation between proteins' chain length, and the average of ΔΔ*G* values of their single nucleotide mutations and all possible mutations. Open circles (○) and broken line are for single nucleotide mutations of natural proteins, filled circles (●) and continuous line are for all possible mutations of natural proteins and filled triangle (▲) is all possible mutations of novel proteins.

by the distribution of computationally designed proteins, a much more "robust" distribution, by which, the deleterious effects of mutations are significantly minimized, and the fraction of stabilizing mutations is larger, is possible (Figure 6). Third, the individual distributions of core and surface are largely Gaussian (Figure 4). The latter two points imply that, although the effects of mutations, whether stabilizing, or destabilizing, are statistically distributed around a certain mean, the mean value might be affected by how a protein was designed, or evolved. It should also be noted that we have analyzed only globular, monomeric, single domain proteins, and many other proteins such as membrane proteins, fibril proteins, or oligomeric proteins, may possess different distributions.

## The mutational robustness of proteins

The tolerance of proteins to mutations is an extensively studied topic. Whilst mutational robustness is not the central topic of this work, our results do relate to certain of its aspects. Experimental measurements of mutational tolerance indicate large variability between proteins. In contrast, the FoldX computations presented here predict that many proteins have a strikingly similar ΔΔ*G* distribution. This discrepancy might be due to several reasons.

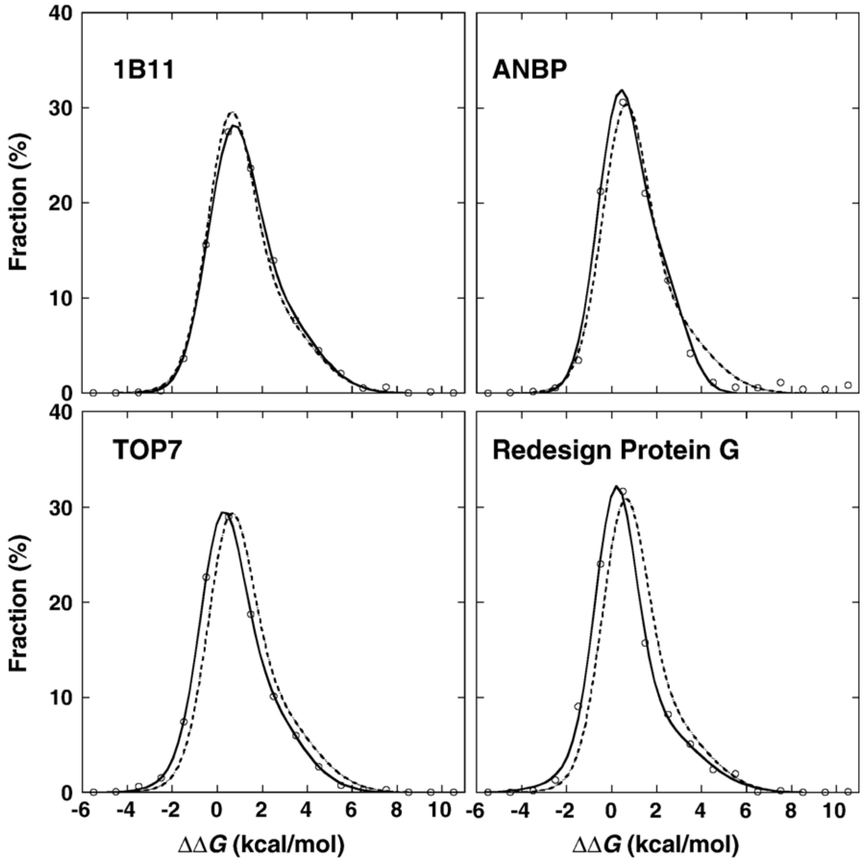

**Figure 6.** The ΔΔ*G* distributions of novel proteins (for details see Table 1). The ΔΔ*G* values were computed by FoldX, and are presented in a histogram as above (Figure 1). The distributions were fitted to a bi-Gaussian function (equation (1); broken dashed line), or to the universal model (equations (2′) and (4′); continuous line). The fits by both these models largely overlap in the case of *in vitro* selected proteins (1B11, ANBP), but differ for the computationally designed TOP7 and redesign protein G.

Biases in the FoldX predictions cannot be ruled out, but as shown above these do not seem to dominate the distributions. In addition, FoldX computes stability effects, but ignores effects on other crucial parameters such as function. Nevertheless, the vast majority of randomly acquired mutations affect stability, and thereby the levels of soluble, active protein.[20,61,62] To our view, there are two other, more likely reasons for this discrepancy. First, experimental measurements of mutational robustness, or neutrality, were performed under very different conditions, and models enabling the quantitative description of robustness have only been recently developed.[7,8,63] Second, we suggest that the mutational robustness of proteins is comprised of two separate components.[8] One component is a threshold of initial stability, which buffers many of the destabilizing effects of the first mutations. Once more mutations accumulate, the excess stability conferred by this threshold is exhausted, and the protein's fitness (expression, *in vivo* stability, and activity levels) declines concomitantly with the decrease in its thermodynamic stability (gradient phase). The threshold correlates primarily with thermodynamic stability (Δ*G* folding). Since the majority of mutations are either neutral or only weakly destabilizing (Figure 1), most of the firstly accumulating mutations would be buffered by this threshold. But, although these mutations may have no immediate effect of the protein's fitness, they do compromise its stability. Thus, as additional mutations accumulate, their effect on fitness will become fully pronounced.[8,64,65] The distribution of ΔΔ*G* effects affects both the threshold and the gradient, and appears to be similar in all natural proteins examined here. What is likely to differ to a much greater extent is the thermodynamic stability, or threshold levels, of these proteins. Previous experimental measurements of mutational robustness did not distinguish between these two components. They typically measured the effects of one, or few mutations, and therefore primarily measured threshold robustness, thus indicating a large variability from one protein to another.

Another issue relates to the question how favorable the distributions of natural proteins are in terms of mutational robustness?[25,28,66] Newly emerging models,[7,8,63] and the ΔΔ*G* distributions presented here, provide a novel quantitative measure of robustness, and a means of computing and comparing the degree of mutational robustness of different proteins. Several interesting conclusions can be derived even from the small set of proteins analyzed here. First, the distribution of novel proteins, selected in the laboratory by only several rounds of mutation and selection, appears to be identical to the distributions of proteins that had been under natural selection for many millions of years (Figure 6). Second, as evident by the distribution of computationally designed proteins, a much more robust distribution, by which, the destabilizing effects of mutations are significantly minimized, and the fraction of stabilizing mutations is larger, is possible (Figure 6). These observations imply (but, by all means, do not prove, or directly indicate) that the stability effects of mutations may not be shaped, or strongly biased, by natural selection. Future research might reveal whether the distributions of certain natural proteins are more robust than those of the average, proteins described here, and whether robust distributions relate to certain evolutionary histories, physiological roles, or organismal features.

Finally, another interesting aspect regards the relationship between protein size and mutational robustness. Our results indicate that the effects of mutations are, on average, less destabilizing in small proteins (Figure 5(d)). This correlation is in agreement with the accepted notion that core residues are more sensitive to mutations than surface residues (Figure 4),[20,24,60] and that smaller proteins have a smaller fraction of core residues (Figure 5(a)). Taken to an extreme, this correlation would indicate that very small proteins with no core would have no strongly destabilizing mutations, but having no core would also imply no defined globular structure. If smaller proteins are more tolerable to mutations, they might also evolve faster. However, a recent study indicated larger proteins, that have a larger fraction of highly contacted residues, evolve faster. This study also noted that, larger proteins exhibit high "designability", which may offset their higher fraction of core residues that are less tolerable to mutations, and hence more slowly evolving.[62] It therefore appears that, whether, and how, size, robustness, and evolvability, correlate is yet an open issue.

## Concluding remarks

The application of FoldX, and possibly of other algorithms that compute the $\Delta\Delta G$ effects of mutations,[30,32–39] towards the prediction of $\Delta\Delta G$ distributions, and the quantitative description of such distributions along the lines described here, are of general utility. The $\Delta\Delta G$ distributions of protein models, including lattice models, are amply generated.[7,25,29] The distributions described here, which are based on force field computations of real proteins validated by experimental data, could be valuable in validating these models, and scaling them to realistic values. Subjected to the caveats described above, the predicted FoldX distributions also provide a quantitative measure of mutational robustness that could be applied towards the comparison of various proteins. Other potential applications include protein design, and in particular, the design of more robust proteins. Foremost, these distributions indicate that key properties of proteins could be explained and predicted by a relatively simple set of rules.

# Methods

## Optimizing models using the FoldX repair function

3D structures were taken from the Protein Data Bank (PDB accession codes are listed in Table 1), and subjected to an optimization procedure using the repair function of FoldX. During this procedure, FoldX identifies the residues that have poor torsion angles, exhibit van der Waals clashes, or total energies. FoldX operates as follows: first, it mutates the selected position to alanine and annotates the side-chain energies of the neighboring residues. Then it mutates the alanine to the selected amino acid, and re-calculates the side-chain energies of the same neighboring residues. Those residues that exhibit an energy difference are then mutated to themselves, to examine if an alternative rotamer will be more favorable. This procedure contains an additional function, where all side-chains are moved slightly in order to eliminate small steric clashes, the value of the steric clash is put at 15 kcal/mol. This quickly eliminates small local clashes, and saves computing time by decreasing the number of valid rotamer searches.

## Generating mutant structures

The mutant structures were generated using the repair position function in FoldX. During this design procedure, FoldX is testing different rotamers and allows neighbor side-chains to move. The program first introduces a mutation to alanine, and then mutates it into the desired residue (while moving the neighbor residues).

## Energy calculations

Energy calculations of mutant proteins were performed with the FoldX energy function that includes terms that have been found to be important for protein stability, where the energy of unfolding ($\Delta G$) of a target protein is calculated using equation (5):

$$\begin{aligned}\Delta G = {} & \Delta G_{\mathrm{vdw}} + \Delta G_{\mathrm{solvH}} + \Delta G_{\mathrm{solvP}} + \Delta G_{\mathrm{wb}} + \Delta G_{\mathrm{Hbond}} \\ & + \Delta G_{\mathrm{el}} + \Delta G_{\mathrm{kon}} + T\Delta S_{\mathrm{mc}} + T\Delta S_{\mathrm{sc}} + T\Delta S_{\mathrm{tr}}\end{aligned} \quad (5)$$

where $\Delta G_{\mathrm{vdw}}$ is the sum of the van der Waals contributions of all atoms, with respect to the same interactions with the solvent; $\Delta G_{\mathrm{solvH}}$ and $\Delta G_{\mathrm{solvP}}$ are the differences in solvation energy for apolar and polar groups, respectively, when going from the unfolded to the folded state; $\Delta G_{\mathrm{Hbond}}$ is the free energy difference between the formation of an intramolecular hydrogen bond compared to intermolecular hydrogen bond formation (with solvent); $\Delta G_{\mathrm{wb}}$, is the extra stabilizing free energy provided by a water molecule making more than one hydrogen bond to the protein (water bridges) that cannot be taken into account with non-explicit solvent approximations; $\Delta G_{\mathrm{el}}$ is the electrostatic contribution of charged groups, including the helix dipole; $\Delta G_{\mathrm{kon}}$ reflects the effect of electrostatic interactions on the $k_{\mathrm{on}}$. $\Delta S_{\mathrm{mc}}$ is the entropy cost for fixing the backbone in the folded state. This term is dependent on the intrinsic tendency of a particular amino acid to adopt certain dihedral angles; $\Delta S_{sc}$ is the entropic cost of fixing a side-chain in a particular conformation ($\Delta S_{sc}$ is the loss of translational and rotational entropy upon making the complex). The energy values of $\Delta G_{vdw}$, $\Delta G_{solvH}$, $\Delta G_{solvP}$ and $\Delta G_{Hbond}$ attributed to each atom type were derived from a set of experimental data, and $\Delta S_{mc}$ and $\Delta S_{mc}$ have been taken from theoretical estimates. The van der Waals contributions are derived from vapor to water energy transfer, while in the protein we are going from solvent to protein. It should be noted that the energy value of van der Waals clash is capped at 1.3 kcal/mol to avoid overestimation of the clash that could be avoidable by backbone relaxation in a real protein structure instead of 15 kcal/mol. The energy values obtained by FoldX were converted to realistic values based on a normalization function obtained by fitting the experimental and computed data (Supplementary Data Figure 1; $\Delta\Delta G^{experiment} = (\Delta\Delta G^{FoldX} + 0.078)/1.14$).

### Data processing

The ASA of each amino acid residue was calculated by the web server program ASA view‡. The $\Delta\Delta G$ values obtained by FoldX were classified to 25 bins, each 1.0 kcal/mol wide, from −10 kcal/mol to 15 kcal/mol (all possible mutations with $\Delta\Delta G > 14$ kcal/mol were classified into the 14–15 kcal/mol bin, and mutations with $\Delta\Delta G < -9$ kcal/mol into the (−10)–(−9) bin). The number of mutations in each bin was counted to make the distribution of $\Delta\Delta G$. Data fitting was performed with KaleidaGraph.

## Acknowledgements

N. T. is a recipient of an EMBO Short Term Fellowship. Financial Support by the Israel Science Foundation is gratefully acknowledged. We are grateful to Shalev Itzkovitz for his invaluable assistance regarding the processing of $\Delta\Delta G$ distributions.

## Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2007.03.069

## References

1. Branden, C. & Tooze, J. (1999). *Introduction to Protein Structure.* Garland, New York.
2. Voigt, C. A., Kauffman, S. & Wang, Z. G. (2000). Rational evolutionary design: the theory of in vitro protein evolution. *Advan. Protein Chem.* **55**, 79–160.
3. Lehmann, M., Pasamontes, L., Lassen, S. F. & Wyss, M. (2000). The consensus concept for thermostability engineering of proteins. *Biochim. Biophys. Acta*, **1543**, 408–415.
4. van den Burg, B. & Eijsink, V. G. (2002). Selection of mutations for increased protein stability. *Curr. Opin. Biotechnol.* **13**, 333–337.
5. DePristo, M. A., Weinreich, D. M. & Hartl, D. L. (2005). Missense meanderings in sequence space: a biophysical view of protein evolution. *Nature Rev. Genet.* **6**, 678–687.
6. Pal, C., Papp, B. & Lercher, M. J. (2006). An integrated view of protein evolution. *Nature Rev. Genet.* **7**, 337–348.
7. Bloom, J. D., Silberg, J. J., Wilke, C. O., Drummond, D. A., Adami, C. & Arnold, F. H. (2005). Thermodynamic prediction of protein neutrality. *Proc. Natl Acad. Sci. USA*, **102**, 606–611.
8. Bershtein, S., Segal, M., Bekerman, R., Tokuriki, N. & Tawfik, D. S. (2006). Robustness-epistasis link shapes the fitness landscape of a randomly drifting protein. *Nature*, **444**, 929–932.
9. England, J. L., Shakhnovich, B. E. & Shakhnovich, E. I. (2003). Natural selection of more designable folds: a mechanism for thermophilic adaptation. *Proc. Natl Acad. Sci. USA*, **100**, 8727–8731.
10. Govindarajan, S. & Goldstein, R. A. (1997). Evolution of model proteins on a foldability landscape. *Proteins: Struct. Funct. Genet.* **29**, 461–466.
11. Bornberg-Bauer, E. & Chan, H. S. (1999). Modeling evolutionary landscapes: mutational stability, topology, and superfunnels in sequence space. *Proc. Natl Acad. Sci. USA*, **96**, 10689–10694.
12. Butterfoss, G. L. & Kuhlman, B. (2006). Computer-based design of novel protein structures. *Annu. Rev. Biophys. Biomol. Struct.* **35**, 49–65.
13. Shortle, D., Stites, W. E. & Meeker, A. K. (1990). Contributions of the large hydrophobic amino acids to the stability of staphylococcal nuclease. *Biochemistry*, **29**, 8033–8041.
14. Green, S. M., Meeker, A. K. & Shortle, D. (1992). Contributions of the polar, uncharged amino acids to the stability of staphylococcal nuclease: evidence for mutational effects on the free energy of the denatured state. *Biochemistry*, **31**, 5717–5728.
15. Meeker, A. K., Garcia-Moreno, B. & Shortle, D. (1996). Contributions of the ionizable amino acids to the stability of staphylococcal nuclease. *Biochemistry*, **35**, 6443–6449.
16. Chen, J. & Stites, W. E. (2001). Energetics of side chain packing in staphylococcal nuclease assessed by systematic double mutant cycles. *Biochemistry*, **40**, 14004–14011.
17. Holder, J. B., Bennett, A. F., Chen, J., Spencer, D. S., Byrne, M. P. & Stites, W. E. (2001). Energetics of side chain packing in staphylococcal nuclease assessed by exchange of valines, isoleucines, and leucines. *Biochemistry*, **40**, 13998–14003.
18. Serrano, L., Kellis, J. T., Jr., Cann, P., Matouschek, A. & Fersht, A. R. (1992). The folding of an enzyme. II. Substructure of barnase and the contribution of different interactions to protein stability. *J. Mol. Biol.* **224**, 783–804.
19. Serrano, L., Day, A. G. & Fersht, A. R. (1993). Step-wise mutation of barnase to binase. A procedure for engineering increased stability of proteins and an experimental analysis of the evolution of protein stability. *J. Mol. Biol.* **233**, 305–312.
20. Matthews, B. W. (1993). Structural and genetic analysis of protein stability. *Annu. Rev. Biochem.* **62**, 139–160.
21. Liu, R., Baase, W. A. & Matthews, B. W. (2000). The

‡ http://www.netasa.org/asaview/

introduction of strain and its effects on the structure and stability of T4 lysozyme. *J. Mol. Biol.* **295**, 127–145.

22. Silverman, J. A., Balakrishnan, R. & Harbury, P. B. (2001). Reverse engineering the (beta/alpha )8 barrel fold. *Proc. Natl Acad. Sci. USA*, **98**, 3092–3097.
23. Kunichika, K., Hashimoto, Y. & Imoto, T. (2002). Robustness of hen lysozyme monitored by random mutations. *Protein Eng.* **15**, 805–809.
24. Cordes, M. H. & Sauer, R. T. (1999). Tolerance of a protein to multiple polar-to-hydrophobic surface substitutions. *Protein Sci.* **8**, 318–325.
25. Taverna, D. M. & Goldstein, R. A. (2002). Why are proteins so robust to site mutations? *J. Mol. Biol.* **315**, 479–484.
26. Guo, H. H., Choe, J. & Loeb, L. A. (2004). Protein tolerance to random amino acid change. *Proc. Natl Acad. Sci. USA*, **101**, 9205–9210.
27. Reddy, B. V., Datta, S. & Tiwari, S. (1998). Use of propensities of amino acids to the local structural environments to understand effect of substitution mutations on protein stability. *Protein Eng.* **11**, 1137–1145.
28. Wagner, A. (2005). *Robustness and Evolvability in Living Systems.* Prinston University Press.
29. Tiana, G., Broglia, R. A. & Provasi, D. (2001). Designability of lattice model heteropolymers. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **64**, 011904.
30. Guerois, R., Nielsen, J. E. & 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.
31. Schymkowitz, J. W., Rousseau, F., Martins, I. C., Ferkinghoff-Borg, J., Stricher, F. & Serrano, L. (2005). Prediction of water and metal binding sites and their affinities by using the Fold-X force field. *Proc. Natl Acad. Sci. USA*, **102**, 10147–10152.
32. Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F. & Serrano, L. (2005). The FoldX web server: an online force field. *Nucl. Acids Res.* **33**, W382–W388.
33. Zhou, H. & 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.
34. Cheng, J., Randall, A. & Baldi, P. (2006). Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins: Struct. Funct. Genet.* **62**, 1125–1132.
35. Gilis, D. & Rooman, M. (2000). PoPMuSiC, an algorithm for predicting protein mutant stability changes: application to prion proteins. *Protein Eng.* **13**, 849–856.
36. Kwasigroch, J. M., Gilis, D., Dehouck, Y. & Rooman, M. (2002). PoPMuSiC, rationally designing point mutations in protein structures. *Bioinformatics*, **18**, 1701–1702.
37. Saunders, C. T. & Baker, D. (2002). Evaluation of structural and evolutionary contributions to deleterious mutation prediction. *J. Mol. Biol.* **322**, 891–901.
38. Parthiban, V., Gromiha, M. M., Hoppe, C. & Schomburg, D. (2007). Structural analysis and prediction of protein mutant stability using distance and torsion potentials: role of secondary structure and solvent accessibility. *Proteins: Struct. Funct. Genet.* **66**, 41–52.
39. Parthiban, V., Gromiha, M. M. & Schomburg, D. (2006). CUPSAT: prediction of protein stability upon point mutations. *Nucl. Acids Res.* **34**, W239–W242.
40. van der Sloot, A. M., Tur, V., Szegezdi, E., Mullally, M. M., Cool, R. H., Samali, A. *et al.* (2006). Designed tumor necrosis factor-related apoptosis-inducing ligand variants initiating apoptosis exclusively via the DR5 receptor. *Proc. Natl Acad. Sci. USA*, **103**, 8634–8639.
41. van der Sloot, A. M., Mullally, M. M., Fernandez-Ballester, G., Serrano, L. & Quax, W. J. (2004). Stabilization of TRAIL, an all-beta-sheet multimeric protein, using computational redesign. *Protein Eng. Des. Sel.* **17**, 673–680.
42. Kiel, C., Wohlgemuth, S., Rousseau, F., Schymkowitz, J., Ferkinghoff-Borg, J., Wittinghofer, F. & Serrano, L. (2005). Recognizing and defining true Ras binding domains II: in silico prediction based on homology modelling and energy calculations. *J. Mol. Biol.* **348**, 759–775.
43. Reichmann, D., Cohen, M., Abramovich, R., Dym, O., Lim, D., Strynadka, N. C. & Schreiber, G. (2007). Binding hot spots in the TEM1-BLIP interface in light of its modular architecture. *J. Mol. Biol.* **365**, 663–679.
44. Kiel, C. & Serrano, L. (2006). The ubiquitin domain superfold: structure-based sequence alignments and characterization of binding epitopes. *J. Mol. Biol.* **355**, 821–844.
45. Aharoni, A., Gaidukov, L., Khersonsky, O., Mc, Q. G. S., Roodveldt, C. & Tawfik, D. S. (2005). The 'evolvability' of promiscuous protein functions. *Nature Genet.* **37**, 73–76.
46. Reetz, M. T. (2004). Changing the enantioselectivity of enzymes by directed evolution. *Methods Enzymol.* **388**, 238–256.
47. Lin, L., Pinker, R. J., Phillips, G. N. & Kallenbach, N. R. (1994). Stabilization of myoglobin by multiple alanine substitutions in helical positions. *Protein Sci.* **3**, 1430–1435.
48. Stefani, M., Taddei, N. & Ramponi, G. (1997). Insights into acylphosphatase structure and catalytic mechanism. *Cell. Mol. Life Sci.* **53**, 141–151.
49. Pickart, C. M. & Eddins, M. J. (2004). Ubiquitin: structures, functions, mechanisms. *Biochim. Biophys. Acta*, **1695**, 55–72.
50. Graur, D. & Li, W. H. (1999). *Fundamentals of Molecular Evolution* (Edition, S., ed), Sinauer Associates, Inc., Massachusetts.
51. Kamtekar, S., Schiffer, J. M., Xiong, H., Babik, J. M. & Hecht, M. H. (1993). Protein design by binary patterning of polar and nonpolar amino acids. *Science*, **262**, 1680–1685.
52. Lee, B. & Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **55**, 379–400.
53. Keefe, A. D. & Szostak, J. W. (2001). Functional proteins from a random-sequence library. *Nature*, **410**, 715–718.
54. Lo Surdo, P., Walsh, M. A. & Sollazzo, M. (2004). A novel ADP- and zinc-binding fold from function-directed in vitro evolution. *Nature Struct. Mol. Biol.* **11**, 382–383.
55. Riechmann, L. & Winter, G. (2000). Novel folded protein domains generated by combinatorial shuffling of polypeptide segments. *Proc. Natl Acad. Sci. USA*, **97**, 10068–10073.
56. Kohl, A., Binz, H. K., Forrer, P., Stumpp, M. T., Pluckthun, A. & Grutter, M. G. (2003). Designed to be stable: crystal structure of a consensus ankyrin repeat protein. *Proc. Natl Acad. Sci. USA*, **100**, 1700–1705.
57. Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L. & Baker, D. (2003). Design of a novel

globular protein fold with atomic-level accuracy. *Science*, **302**, 1364–1368.

58. Nauli, S., Kuhlman, B., Le Trong, I., Stenkamp, R. E., Teller, D. & Baker, D. (2002). Crystal structures and increased stabilization of the protein G variants with switched folding pathways NuG1 and NuG2. *Protein Sci.* **11**, 2924–2931.
59. Nauli, S., Kuhlman, B. & Baker, D. (2001). Computer-based redesign of a protein folding pathway. *Nature Struct. Biol.* **8**, 602–605.
60. Bowie, J. U., Reidhaar-Olson, J. F., Lim, W. A. & Sauer, R. T. (1990). Deciphering the message in protein sequences: tolerance to amino acid substitutions. *Science*, **247**, 1306–1310.
61. Godoy-Ruiz, R., Perez-Jimenez, R., Ibarra-Molero, B. & Sanchez-Ruiz, J. M. (2004). Relation between protein stability, evolution and structure, as probed by carboxylic acid mutations. *J. Mol. Biol.* **336**, 313–318.
62. Bloom, J. D., Drummond, D. A., Arnold, F. H. & Wilke, C. O. (2006). Structural determinants of the rate of protein evolution in yeast. *Mol. Biol. Evol.* **23**, 1751–1761.
63. Bloom, J. D., Raval, A. & Wilke, C. O. (2007). Thermodynamics of neutral protein evolution. *Genetics*, **175**, 255–266.
64. Bloom, J. D., Labthavikul, S. T., Otey, C. R. & Arnold, F. H. (2006). Protein stability promotes evolvability. *Proc. Natl Acad. Sci. USA*, **103**, 5869–5874.
65. Besenmatter, W., Kast, P. & Hilvert, D. (2007). Relative tolerance of mesostable and thermostable protein homologs to extensive mutation. *Proteins: Struct. Funct. Genet.* **66**, 500–506.
66. de Visser, J. A., Hermisson, J., Wagner, G. P., Ancel Meyers, L., Bagheri-Chaichian, H., Blanchard, J. L. *et al.* (2003). Perspective: evolution and detection of genetic robustness. *Evol. Int. J. Org. Evol.* **57**, 1959–1972.

***Edited by B. Honig***

*(Received 31 October 2006; received in revised form 22 March 2007; accepted 27 March 2007)*
Available online 31 March 2007