JMB

Can Contact Potentials Reliably Predict Stability of Proteins?

Jainab Khatun, Sagar D. Khare and Nikolay V. Dokholyan*

Department of Biochemistry and Biophysics, University of North Carolina at Chapel Hill Chapel Hill, NC 27599, USA

The simplest approximation of interaction potential between amino acid residues in proteins is the contact potential, which defines the effective free energy of a protein conformation by a set of amino acid contacts formed in this conformation. Finding a contact potential capable of predicting free energies of protein states across a variety of protein families will aid protein folding and engineering *in silico* on a computationally tractable time-scale. We test the ability of contact potentials to accurately and transferably (across various protein families) predict stability changes of proteins upon mutations. We develop a new methodology to determine the contact potentials in proteins from experimental measurements of changes in protein's thermodynamic stabilities ($\Delta\Delta G$) upon mutations. We apply our methodology to derive sets of contact interaction parameters for a hierarchy of interaction models including solvation and multi-body contact parameters. We test how well our models reproduce experimental measurements by statistical tests. We evaluate the maximum accuracy of predictions obtained by using contact potentials and the correlation between parameters derived from different data-sets of experimental ($\Delta\Delta G$) values. We argue that it is impossible to reach experimental accuracy and derive fully transferable contact parameters using the contact models of potentials. However, contact parameters may yield reliable predictions of $\Delta\Delta G$ for datasets of mutations confined to the same amino acid positions in the sequence of a single protein.

© 2004 Elsevier Ltd. All rights reserved.

Keywords: contact potential; protein stability; mutation; protein folding; protein design

*Corresponding author

Introduction

A fundamental goal of molecular biophysics is to understand the relationship between protein sequence and structure, a problem known as the protein folding problem.[1–8] Conversely, identifying amino acid sequences that fold rapidly and are stable in a target conformation, is known as the protein design problem.[9–20] Despite recent remarkable successes in protein folding *in silico*,[21–24] the folding time-scales of most proteins are not accessible to detailed computational techniques, such as molecular mechanics simulations. This limitation makes identification of the folding properties of naturally occurring and designed proteins intractable.

An alternative approach to the solution of both protein folding and design problems is the development of a simplified model of amino acid interactions that can be used in rapid computational techniques, such as Monte–Carlo simulations[25–27] and discrete molecular dynamics.[28–31] Such models incorporate a coarse-grained description of the polypeptide chain where each residue is represented by a limited (one or a few) number of beads connected by a chain. Effective interactions are assigned between beads. The model drastically reduces the computational cost of folding, and is amenable to detailed statistical mechanical characterization.

One of the widely used forms of simplified inter-residue potentials is the contact potential in which amino acids interact if they are located spatially within a certain distance from each other.[31–38] The quasichemical approximation is a method for deriving pairwise contact potentials from the number of residue–residue contacts found in a

J.K. and S.D.K. contributed equally to this work.

Abbreviations used: SVD, singular value decomposition; DS*n*, dataset *n*.

E-mail address of the corresponding author: dokh@med.unc.edu

large set of protein structures.[37] The quasichemical approximation has proved to be successful in designing and folding proteins on lattices.[38] A strategy proposed by Clementi *et al.* to derive pairwise contact potentials for off-lattice protein models was to iteratively refine the contact potentials by starting from a trial set of interaction parameters, and modifying these parameters until the energy minimum conformation corresponded to the native-state of a given protein.[32] Efforts have been made to find contact interaction parameters such that the energies of the native states of a set of proteins are simultaneously smaller than their respective decoy conformations.[33,36] Potentials derived from these methods have successfully captured the main features of the statistical mechanics of proteins but have failed to consistently reach protein native states in folding simulations. The motivation for these studies was to find a contact potential that does not require prerequisite knowledge of protein structures and is capable of folding a model protein to native conformation. On the other hand, the structure-based Gō potential[31,34] has been successful in reaching protein native state in molecular dynamics simulations.[28,39] However, since Gō potential is derived in the context of a specific protein native state, it is not transferable to other proteins. A fully transferable universal potential, therefore, remains elusive.

The observed non-transferability of contact potentials raises the questions of whether it is possible to derive a set of contact interactions that are able to (a) discriminate the native states of a set of proteins from all structural decoys, and (b) capture protein thermodynamics to predict the changes in free energy of unfolding of a set of proteins upon single or multi-point mutations with experimental accuracy? It was shown by Vendruscolo *et al.* that no pairwise contact potential exists that can stabilize an off-lattice model of some proteins relative to their structural decoys.[40,41] The ability of contact models to predict stability changes upon mutations ($\Delta\Delta G$) has not been similarly evaluated. Many successful models for predicting stability changes upon mutations either use detailed representation of the polypeptide chain with continuous potentials[42,43] and/or database-derived statistical potentials.[44]

One approach for deriving interaction parameters for predicting stability changes in proteins upon mutation is to "learn" these parameters by fitting the interaction model to experimental data.[43,45] There are two requirements for a derived

interaction potential for predicting stability changes of proteins upon mutations: such interaction potential predicts $\Delta\Delta G$ (within experimental error) for (i) the set of proteins from which it is derived (the criterion of accuracy), and (ii) sets of proteins that were not used for derivation of this interaction potential (the criterion of transferability). Here, we ask whether simple contact models can be used to predict $\Delta\Delta G$ and evaluate the reliability of contact potentials using the criteria of accuracy and transferability. The contact potentials are parameterized from experimental $\Delta\Delta G$ values using the singular value decomposition (SVD) algorithm. We show that contact potentials fail with respect to both the criteria; i.e. they are neither accurate nor transferable. We argue that it is impossible to determine accurate transferable contact parameters to predict accurately the changes in stability of a set of proteins simultaneously.

Results and Discussion

We derive a hierarchy of amino acid interaction potentials from several datasets of experimentally measured changes in free energy upon mutation (Methods). In order to test accuracy as well as transferability of potentials, we choose three different datasets (Table 1). (i) dataset 1 (DS1) consists of 303 four-point mutants of eglin c, a small 70 residue protease inhibitor structurally homologous to chymotrypsin inhibitor 2. The $\Delta\Delta G$ values for eglin c were determined with great accuracy (up to ± 0.087 kcal/mol (1 cal = 4.184 J)) by Edgell *et al.* for (quadruple) mutations at R22, E23, T26 and L27 in the solvent-exposed helix of eglin c.[46] (ii) DS2 consists of 658 single, double, triple and quadruple mutants for staphylococcal nuclease (149 amino acid residues). We use $\Delta\Delta G$ values reported for staphylococcal nuclease in the ProTherm database[47] and by Zhou *et al.*[48] The typical errors reported for staphylococcal nuclease mutant $\Delta\Delta G$ values are approximately ± 0.1 kcal/mol. The mutation sites are not localized to any particular sub-structure of the protein. (iii) DS3 consists of 1356 $\Delta\Delta G$ measurements for a set of 11 proteins that are reported in ProTherm,[47] and by Guerois *et al.*[43] and Zhou *et al.*[48] The mutation sites are scattered throughout the structure of these 11 proteins. The maximum error bar for this dataset is 0.48 kcal/mol. We derive potentials for each data set and compute the accuracy of our predictions by evaluating the models' predictive capability. We test the ability of

Table 1. Summary of datasets that we use to derive the contact interactions in proteins

Dataset	Protein (PDB code)	No. mutants	Source	References
DS1	Eglin c (1EGL)	303	Edgell *et al.*	46,60
DS2	Staphylococcal nuclease (1STN)	658	ProTherm database, Guerois *et al.*, Zhou *et al.*	43,47,48,79
DS3	1STN, 1ARR, 1BNI, 1BPI, 1BP2, 1BVC, 1C90, 1FKJ, 2ABD, 2LZM, 1EGL	1356	Edgell *et al.*, ProTherm database, Guerois *et al.*, Zhou *et al.*	3,46–48, 60,79

potentials derived from one dataset to predict changes in stability of other datasets (transferability).

Predicting $\Delta\Delta G$ using contact models

Two-body interaction potential

We consider pairwise two-body interaction (H_2, equation (3)) and derive interaction parameters $u_2(\sigma_a,\sigma_b)$ (210 parameters for a 20 letter amino acid alphabet) from a set of equations (such as equation (6)) corresponding to all the mutants in a given dataset, by employing the SVD algorithm. Using derived interaction parameters, we predict the change in free energy ($\Delta\Delta G$) upon mutation and examine our prediction quality by both the resubstitution and the jack-knife tests (see Methods) for all the three datasets.

The resubstitution test is crucial, because it reflects the self-consistency of the prediction algorithm and sets an upper bound on the performance of a given model. When the experimental dataset is limited, the resubstitution test is useful for assessment of the accuracy of an interaction model,[49] while the jack-knife test provides a rigorous assessment of the model's predictive power. In both tests, we compute the linear regression correlation coefficients between the predicted and actual experimental values (Table 2). For the more accurate eglin c measurements (DS1), we obtain resubstitution (r_s) and jack-knife (r_j) correlation coefficients of 0.87 and 0.80, respectively (Figure 1(a)), while the values for the same quantities are 0.79 and 0.52 for DS2 (Figure 1(b)) and 0.66 and 0.46 for DS3 (Figure 1(c)).

The correlation between the predicted and actual values of $\Delta\Delta G$ is expected to decrease in the jack-knife test in comparison with the resubstitution test. For both tests, we observe that there are common outliers for which predicted $\Delta\Delta G$ values are significantly different from the experimental values. Specifically, for DS1, the outliers are the

four four-point mutants (R22K/E23K/T26N/L27K, R22N/E23K/T26N/L27A, R22E/E23G/T26E/L27G and R22P/E23H/T26E/L27K) for which the $\Delta\Delta G$ values are 2.488, 2.623, −5.676 and −5.698. Most experimentally measured $\Delta\Delta G$ values in DS1 lie between approximately −3.0 and 0.5 kcal/mol (the average and the standard deviations of $\Delta\Delta G$ measurements are −1.393 and 0.83, respectively). Such large perturbations of free energy in the outliers indicate large structural alterations that, in turn, imply that the contact approximation is not valid in these cases. If the outliers are ignored, the values of r_s and r_j are 0.93 and 0.89, respectively.

The significantly better performance of the model for DS1 compared to DS2 and DS3 can be attributed, partially, to greater accuracy (error of measurement 0.087 kcal/mol) and greater consistency in experimental measurements (all data points are obtained from one source and at the same experimental conditions). The stability determination of any given protein is a function of the conditions under which the measurements are made. For example, pH and salt concentrations can alter the measured stability of a particular protein dramatically.[50–52]

Apart from the variability of sources of error in experimental data, the statistical properties of the dataset also determine the performance of a model in the jack-knife test. The more a given interaction parameter is represented in the set of linear equations, the less likely is its estimation to significantly change in the jack-knife test. We estimate the representation of parameters in the dataset using the coverage parameter, C_P, which we define as:

$$C_p = \frac{1}{N_P} \sum_{f=0}^{f=1} fn_f \qquad (1)$$

where n_f is the number of parameters occurring

Table 2. The correlation coefficient between predicted and experimental values of $\Delta\Delta G$ for self-consistency (r_s) and jack-knife (r_j) tests using Hamiltonians of various complexities

Protein	No. mutants	Terms included in H	Self-consistency test		Jack-knife test	
			r_s	$\log_{10} p$	r_j	$\log_{10} p$
Eglin c	303	H_2	0.87	−52	0.80	−44
		$H_2 + H_s$	0.87	−52	0.78	−42
		$H_2 + H_s + H_3$	0.87	−52	0.78	−42
		$H_2 + H_s + H_3{}^{a}$	0.87	−52	0.77	−41
Staphylococcal nuclease	658	H_2	0.79	−91	0.52	−39
		$H_2 + H_s$	0.80	−92	0.47	−33
		$H_2 + H_s + H_3$	0.98	< −100	0.28	−12
		$H_2 + H_s + H_3{}^{a}$	0.88	< −100	0.47	−33
Combined	1356	H_2	0.66	< −100	0.46	−64
		$H_2 + H_s$	0.65	< −100	0.45	−62
		$H_2 + H_s + H_3$	0.91	< −100	0.11	−05
		$H_2 + H_s + H_3{}^{a}$	0.69	< −100	0.45	−62

H_2, H_3, and H_s denote two-body, three-body and solvation Hamiltonian interaction terms, respectively. p is the measure of the statistical significance of the correlation coefficient (the probability of observing these correlations by chance).
[a] The three-body interactions are calculated using only six different types of amino acid residue.
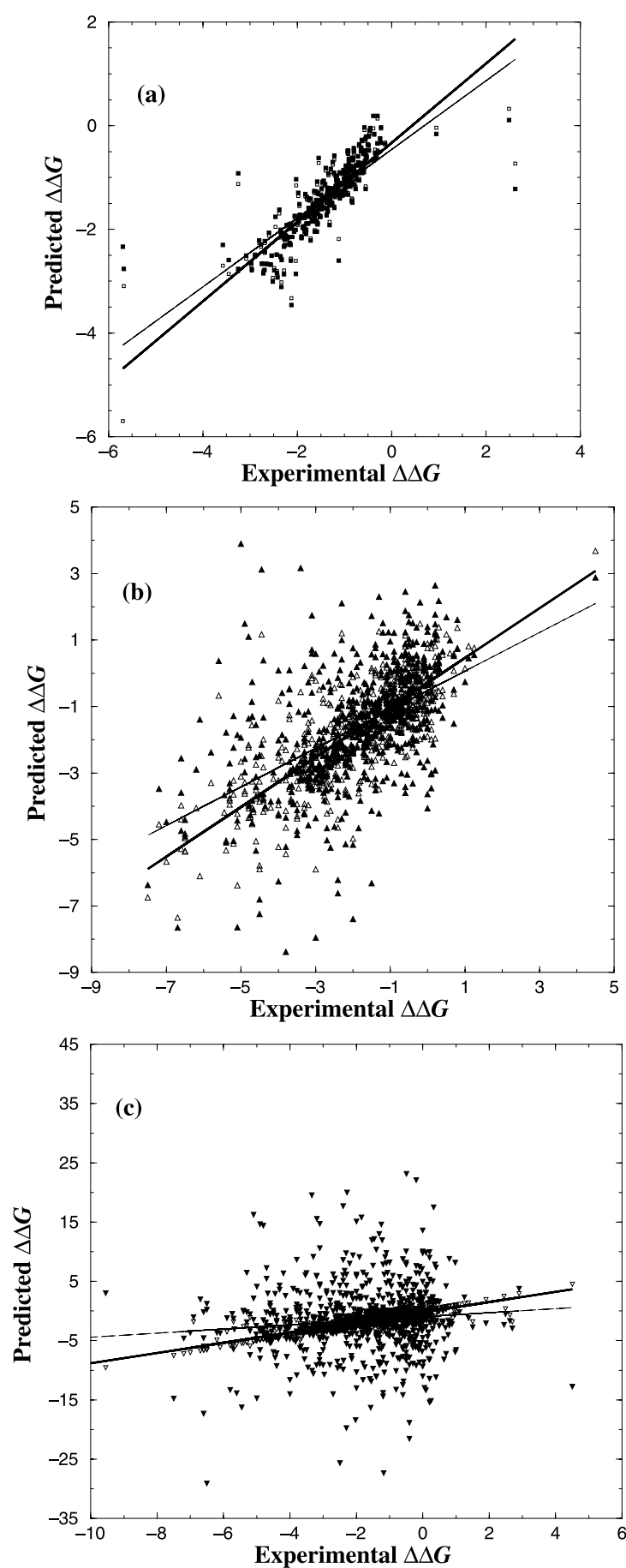
Figure 1. Calculated $\Delta\Delta G$ values compared to the experimental $\Delta\Delta G$ values: (a) for 303 mutants of eglin c (dataset DS1) for self-consistency ($\square$) and jack-knife ($\blacksquare$) tests; (b) for 658 mutants of staphylococcal nuclease (dataset DS2) for self-consistency ($\triangle$) and jack-knife ($\blacktriangle$) tests; and (c) for 1356 mutants of a combined set of 11 proteins (dataset DS3) for self-consistency ($\triangledown$) and jack-knife ($\blacktriangledown$) using two-body Hamiltonian. The two continuous lines represent the linear regression obtained for the self-consistency (thick) and the jack-knife (thin) tests for all the datasets. The correlation coefficient (r) and the probabilities that this correlation is observed at random (p-values) for these tests are reported in Table 2.

with frequency f and N_p is the total number of parameters. For a dataset with all parameters represented in all equations, $C_p = 1$. Small values of C_p indicate that a large number of parameters are under-represented in the set of equations corresponding to the dataset. The C_p values for datasets DS1, DS2 and DS3 are 0.349, 0.045 and 0.044, respectively (Figure 2). Thus, the jack-knife performance is related directly to the parameter space coverage.

In addition, we note that DS1 consists exclusively of mutations localized to four specific positions on eglin c (R22, E23, T26 and L27). DS2 consists of mutations in one protein staphylococcal nuclease, albeit scattered throughout its structure; DS3 consists of a variety of mutations in a variety of proteins. Thus, DS2 and DS3 contact parameters are averaged over diverse structural environments. We find that contact energies derived from mutations that are localized within similar structural environment (e.g. mutants on the solvent-exposed helix in eglin c, DS1) are more successful in self-consistently predicting the respective energetics than contact energies derived from averaging over a large set of environments and conditions (as in DS2 and DS3). The above observation confirms that contact energies are not independent of mutants' structural environments suggesting that including additional terms in the interaction model may capture sufficient environmental detail to successfully predict protein stability.

Adding solvent and multi-body terms

It has been argued that solvation and multi-body effects are important to capture structural environments of amino acid residues.[53-55] Therefore, we add the solvation effect equation (4) in our potential model and derive the interaction with solvent $u_s(\sigma_a)$ along with two-body interaction parameters $u_2(\sigma_a, \sigma_b)$. Following Eisenberg *et al.*,[56] we estimate the solvation energy as a linear function of the solvent-accessible surface area of the wild-type residue. There are a total of 230 (20 solvation + 210 two-body) parameters for the 20-letter amino acid alphabet size. We predict the change in free energy upon mutation and examine the prediction quality of the new derived potential (Table 2). We obtain correlation coefficient for self-consistency and jack-knife tests to be 0.80 and 0.78 for the dataset DS1, 0.80 and 0.47 for DS2, and 0.65 and 0.45 for DS3, respectively. We find that for all databases, the inclusion of the solvation term does not affect either the self-consistency or jack-knife tests performance significantly. An accurate treatment of solvation considering atomistic details of the polypeptide chain may increase the possibility of finding better correlation between experimental and predicted values of free energy changes.

Finally, we add the three-body term H_3 (equation (5)) in our interaction potential model, so that the resulting Hamiltonian consists of the terms H_2
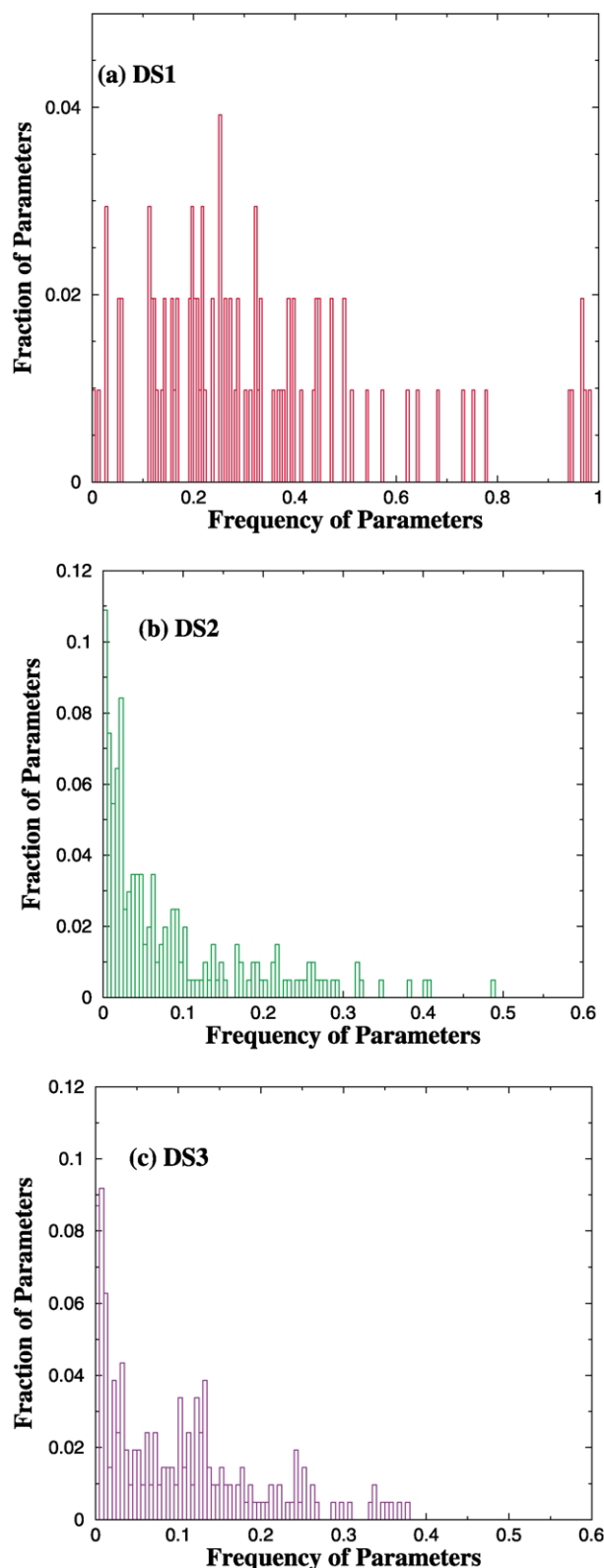

Figure 2. The fraction of occurring parameters as a function of frequencies for the two-body interaction Hamiltonian ($H = H_2$) : for (a) DS1, (b) DS2 and (c) DS3.

(equation (3)), H_s (equation (4)) and H_3 (equation (5)). There are a total of 1770 parameters (210 two-body, 20 solvation and 1540 three-body) to be determined for a 20-letter alphabet size. The

correlation coefficient including three-body interaction for self-consistency and jack-knife tests are 0.87 and 0.78 for DS1, 0.98 and 0.28 for DS2, and 0.91 and 0.11 for DS3, respectively. We find that the three-body Hamiltonian performs better for the self-consistency test but leads to a worse performance on jack-knife tests. This observation can be rationalized as follows: we need to determine 1770 parameters on inclusion of three-body interactions while we have 303, 658 and 1356 equations in DS1, DS2 and DS3, respectively. Therefore, the set of equations corresponding to H_3 represents an underdetermined set with more variables than the number of equations. The parameter space coverage for the $H_2 + H_s + H_3$ Hamiltonian does not change significantly from when we consider only the H_2 Hamiltonian: the C_p values are 0.34, 0.02 and 0.01 for DS1, DS2 and DS3, respectively (Figure 3).

It is possible that the performance of $H_2 + H_s + H_3$ Hamiltonian model is a reflection of the limited number of data points in any dataset used to calculate the energy parameters; consequently, these experimental datasets are not large enough for predicting all 1770 parameters. Sufficient sampling of the parameter space can be obtained by rationally designed, large-scale mutant libraries. While the inclusion of multi-body effects may increase the self-consistency and jack-knife test correlations, we argue below that the solutions may not be universal, i.e. we will not be able to predict stability changes accurately on datasets other than the one from which the parameters are derived.

In order to circumvent the paucity of accurate data, we use a reduced amino acid alphabet to obtain a smaller number of parameters. The minimal amino acid alphabet for protein folding has been an active area of research.[57-59] Recent studies have suggested that the full sequence complexity is not required to design a foldable protein and an alphabet size of six may be sufficient to fold a protein.[57] Thus, we reduce the number of parameters using six types of amino acids (Methods) for the three-body interaction; the two-body and the solvent effects are parameterized for the 20 amino acid alphabet. There are a total of 286 parameters to be determined (210 two-body and 20 solvation terms for 20-letter alphabet, 56 three-body terms for six-letter alphabet size). The performance of this partially reduced alphabet model is not significantly different from the previous model with the full 20-letter amino acid alphabet. The histograms (not shown) of parameter space for this model indicate that the reduction of the alphabet size does not lead to significant improvement in the parameter coverage due to the mutual cancellation of parameter terms in the equations.

Accuracy of parameters derived from contact models

The relatively high correlation coefficients obtained, ($r_s = 0.87$; with H_2 for DS1), indicate that a high proportion of the data can be explained
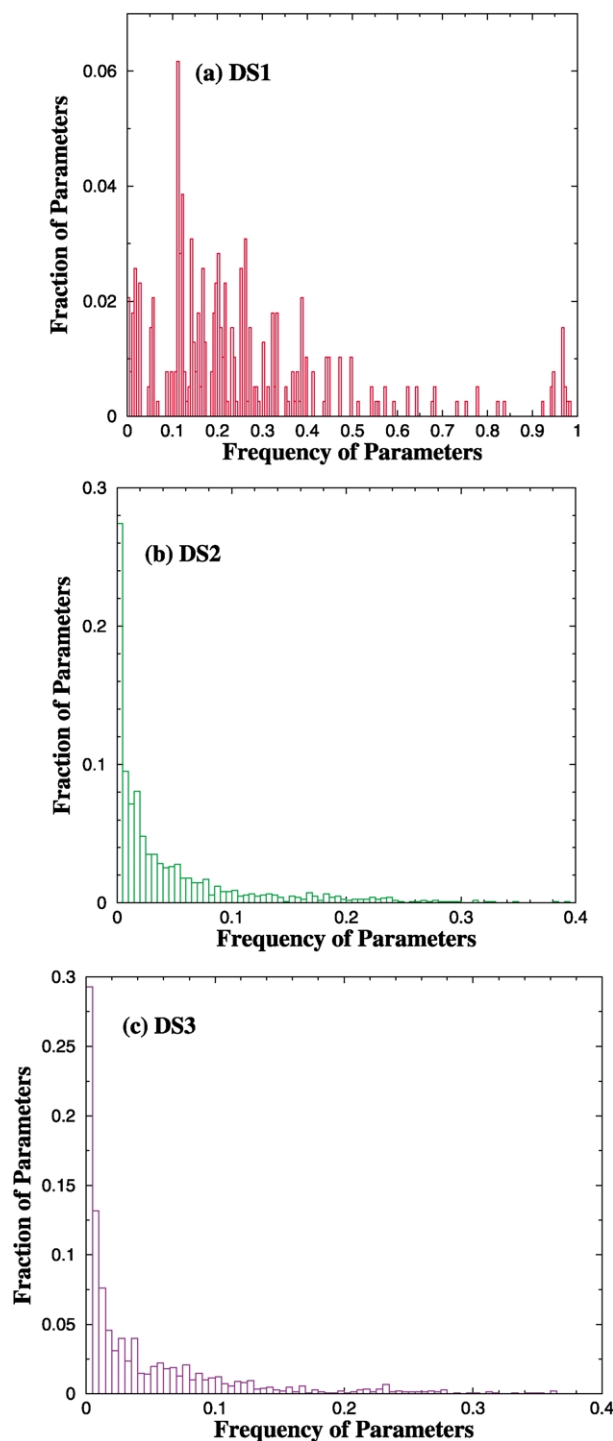

Figure 3. The fraction of parameters as a function of their frequency of occurrence for datasets (a) DS1, (b) DS2 and (c) DS3. The histograms are obtained for interaction models that include two-body, three-body and solvent interactions terms $(H = H_2 + H_s + H_3)$. The comparison of Figures 2 and 3 indicates that the inclusion of three-body interactions reduces the parameter space coverage in the composite Hamiltonian $(H = H_2 + H_s + H_3)$ compared to the two-body Hamiltonian $(H = H_2)$ parameter coverage.

by our model. However, the measure of a model's predictive capability is the uncertainty in the predictions compared to the experimental errors. Therefore, we analyze the accuracy of $\Delta\Delta G$ predictions and compare this accuracy with the experimental uncertainty.

As described in Methods, we solve a set of linear equations $A\mathbf{x} - \mathbf{b} = 0$, where A corresponds to the matrix of contact interactions, $\mathbf{x}$ is the solution vector of contact energies and $\mathbf{b}$ is the experimentally determined $\Delta\Delta G$ values. Using the SVD algorithm, we find the linear least-squares solution, the solution vector for which:

$$\chi^2 = \|(A\mathbf{x} - \mathbf{b})\|_2 / N_e$$

is minimal (χ^2_{min}), where N_e is the number of experimental measurements. χ^2 depends on a tunable threshold parameter ε, which determines the number of non-zero singular eigenvalues of the matrix A. For ε in the range 10^{-3}–10^{-14}, the χ^2 is constant and minimal (see Figure 4). Therefore, we use a value of ε corresponding to the above range ($\varepsilon = 10^{-8}$). With this value of ε, we obtain the solution vector and the associated χ^2_{min}. The value of χ^2_{min} is the parameter that represents the average uncertainty in our prediction, and is therefore the relevant parameter to compare with experimental uncertainties in measurement.

For all the datasets, we find that the uncertainty in our predictions (χ^2_{min}) is greater than the observed experimental error. Thus, we can never reach $\chi^2_{min} \leq \chi^2_{exp}$, where χ^2_{exp} is the square of the experimental error-bar. The χ^2_{min} for dataset DS1 (using only two-body Hamiltonian) is approximately 0.40 kcal/mol, while the experimental error is 0.087 kcal/mol (Figure 4(a)).[46,60] For the dataset DS2, we obtain χ^2_{min} to be equal to 0.9 kcal/mol (Figure 4(b)), while the experimental error is approximately 0.1 kcal/mol.[50,61,62] For the dataset DS3, we obtain χ^2_{min} to be equal to 1.31 kcal/mol (Figure 4(c)), while the maximal experimental error is approximately 0.48 kcal/mol.[63] We obtain similar results upon inclusion of the other interaction terms (H_s and H_3) (data not shown). SVD, with an appropriate choice of the ε parameter, is guaranteed to yield the most accurate least-squares solution for a given dataset. Thus, for all datasets, we show that the uncertainties in prediction are always greater than the corresponding experimental error.

Transferability of parameters derived from contact models

In order to derive a universal set of contact interaction parameters, it is crucial that parameters derived from one dataset accurately predict the $\Delta\Delta G$ values for other datasets. Therefore, we analyze the transferability of interaction parameters between the datasets. We determine the correlation coefficients between the parameters derived from different sets using the two-body Hamiltonian and

find that the correlation is negligible (Figure 5). The lack of correlation between parameters that include solvent and the multi-body effects derived from different datasets (data not shown) also demonstrates the non-transferability of interaction parameters.

The uniqueness of the derived solution depends on the relative number of parameters (N_p) and number of experimental observations (N_e). There are three possibilities to consider for evaluating the uniqueness of a set of contact parameters.

Case 1: $N_e < N_p$. This case represents an underdetermined set of equations, so that we always find more than one least-squares solution. From among these different solutions we choose the minimal norm solution vector corresponding to χ^2_{min}.

Case 2: $N_e > N_p$. This case represents an overdetermined set of equations so that no exact solution exists. We find the least-squares solution that is the closest to satisfying all the equations simultaneously (by minimizing χ^2). However, we find that the number of non-zero singular eigenvalues and, hence, the rank of the matrix A (N_k) is less than the number of parameters (N_p) to be determined. Thus, there is more than one set of solution vectors $\mathbf{x}$ that correspond to the value of χ^2_{min}. The number of least-squares solution vectors is determined by the eigenvalue threshold ε (Figure 6). When $\varepsilon = 0$, which corresponds to considering all the singular values in the SVD procedure ($N_k = N_p$), we obtain a unique solution. As we increase the value of ε, N_k becomes smaller than N_p, and we obtain a number of solutions. If each choice of N_k parameters for a given value of N_p corresponds to a different solution compatible with a given value of χ^2_{min}, then we estimate the number of least-squares solutions as the number of ways of choosing N_k parameters out of N_p. When $\varepsilon = 1$, $N_k = 1$, corresponding to solving one equation, and the number of solutions is N_p.

Case 3: $N_e = N_p$. This case results in a unique solution if the set of equations is not redundant. However, for our datasets, the solution depends on which measurements we choose to derive the parameters from, so that we obtain different solutions depending on the choice of mutants. If the equations are redundant, then finding the least-squares solution reduces to case 1 and there is more than one solution in this case.

For all the above cases 1–3, we obtain more than one least-squares solution. We summarize the balance between uniqueness and accuracy of the derived solutions in a cartoon representation in Figure 7. As stated earlier, the threshold value ε determines the number of non-zero singular values and, hence, both the accuracy and the number of solutions. For a specific value of ε ($\varepsilon = \varepsilon_{min}$), the value of χ^2 is minimal. For $\varepsilon < \varepsilon_{min}$, we include smaller singular values in the solution, so that the components of the solution vector corresponding to these singular values are error-prone (see Methods) and yield $\chi^2 > \chi^2_{min}$. For $\varepsilon > \varepsilon_{min}$, we
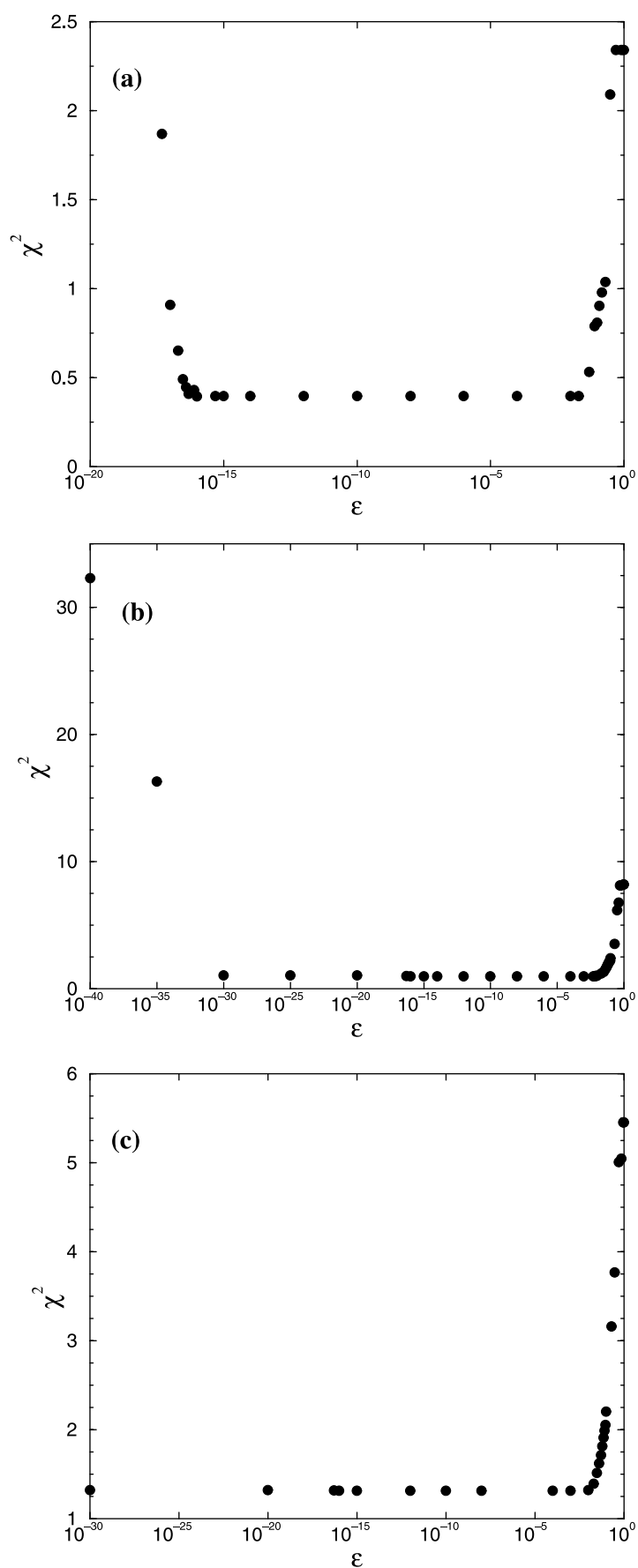
Figure 4. The dependence of χ^2 on ε for datasets (a) DS1, (b) DS2, and (c) DS3. The χ^2 values remain constant for a range of ε (10^{-3} to 10^{-14}) and increase for values of ε outside this range. For DS3, the increase for smaller values of ε is less pronounced. For each dataset, we calculate the linear least-squares solution at $\varepsilon = 10^{-8}$ and obtain the minimal χ^2 value.
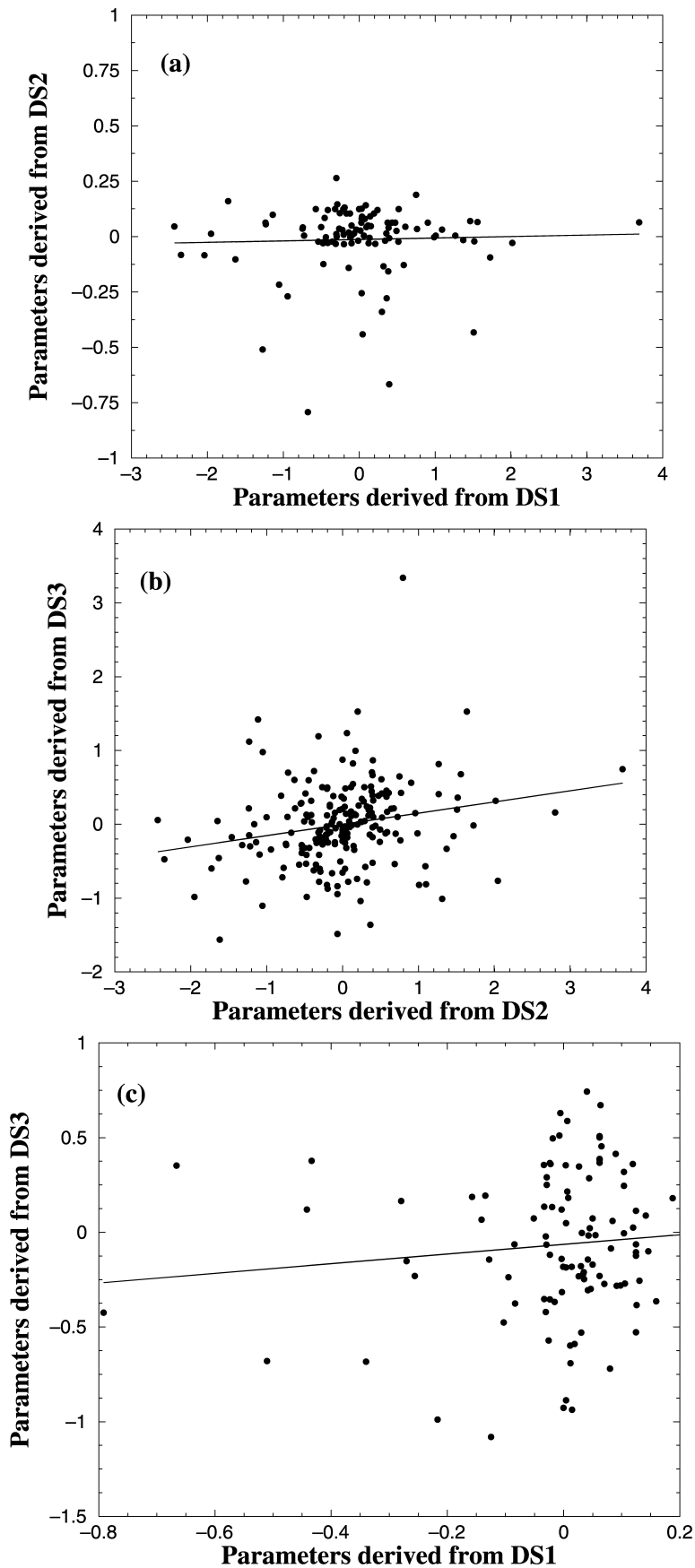
Figure 5. The comparison of the interaction parameters derived from datasets (a) DS1 and DS2, (b) DS2 and DS3, and (c) DS1 and DS3. The interaction parameters for each individual dataset DS1, DS2, and DS3 are obtained by minimizing simultaneously the χ^2 and the length of the solution vector (see Methods). The continuous line represents the linear regression between the parameters derived from different datasets. The linear regression correlation coefficient between parameters derived from datasets DS1 and DS2, DS2 and DS3, and DS1 and DS3 is 0.06, 0.15 and 0.21, respectively.
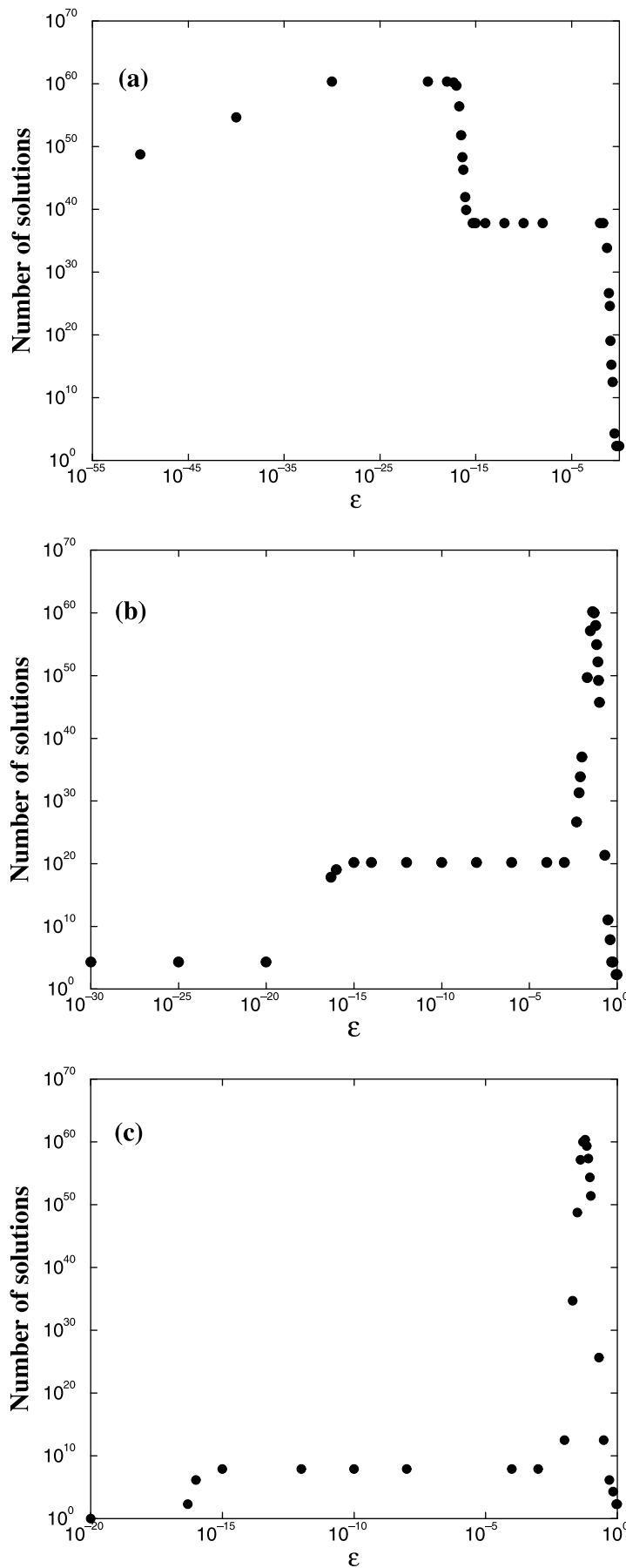
Figure 6. The number of solutions as a function of the threshold value ε for datasets (a) DS1, (b) DS2, and (c) DS3. For all datasets, at $\varepsilon = 0$, the rank of the contact matrix A is equal to the total number of parameters, and we obtain a unique solution. As we increase ε, the rank becomes smaller than the number of parameters, and the number of least-squares solutions is much greater than 1. When $\varepsilon = 1$, the set of linear equations is reduced to one equation, and we obtain N_p solutions.

discard some significant singular eigenvalues, and the number of effective parameters is not sufficient to fit the data with experimental accuracy. As a result, we again obtain $\chi^2 > \chi^2_{min}$. Thus, the accuracy of obtained solutions has a minimum at ε_{min}, and increases for all other values of ε. The uniqueness of solutions is also a function of ε. We obtain a unique solution when $\varepsilon = 0$, and N_p solutions when $\varepsilon = 1$ (Figures 6 and 7). However, for both cases, the corresponding values of χ^2 are much larger than χ^2_{min}. For all datasets, when ε is minimal ($\varepsilon = \varepsilon_{min}$), which corresponds to the highest accuracy of prediction, the number of solutions is much greater than one. Therefore, no unique solution exists and we argue that transferability of parameters from one dataset to the other cannot be achieved.

Implications for the assumptions and usefulness of the contact model

Two assumptions implicit in the contact approximation, first, that contact energies are independent of the protein environment and, second, that the unfolded states are unstructured and not affected by mutation, may be violated, leading to the observed results. We find that when contact energies are derived using studies performed under the same conditions (e.g. temperature, pH, and salt concentration) for mutations at the same residues in a single protein (DS1 for eglin c), we obtain a greater accuracy of prediction in both the self-consistency and jack-knife tests than for heterogeneous DS2 and DS3 datasets. This implies that contact energies are a function of the environment in which a given contact is formed. Thus, the usefulness of contact models is dependent on the experimental database under consideration. For a dataset of mutants corresponding to identical positions in the sequence of a single protein (dataset DS1), contact potentials can reliably predict the stability changes upon mutation (correlation coefficient of 0.79 in the jack-knife test) and are therefore reasonably adequate for prediction of similar mutants in the protein. However, for datasets comprising of either mutants corresponding to several distinct amino acid positions in the sequence of a given protein (dataset DS2) or a set of proteins (dataset DS3) the corresponding correlation coefficients in the jack-knife test are 0.52 and 0.46, indicating their predictive inadequacy.

Furthermore, it has been noted[64-67] that there is residual structure in the unfolded states, and it is likely that effect of mutations on unfolded states may contribute significantly to the protein stability.
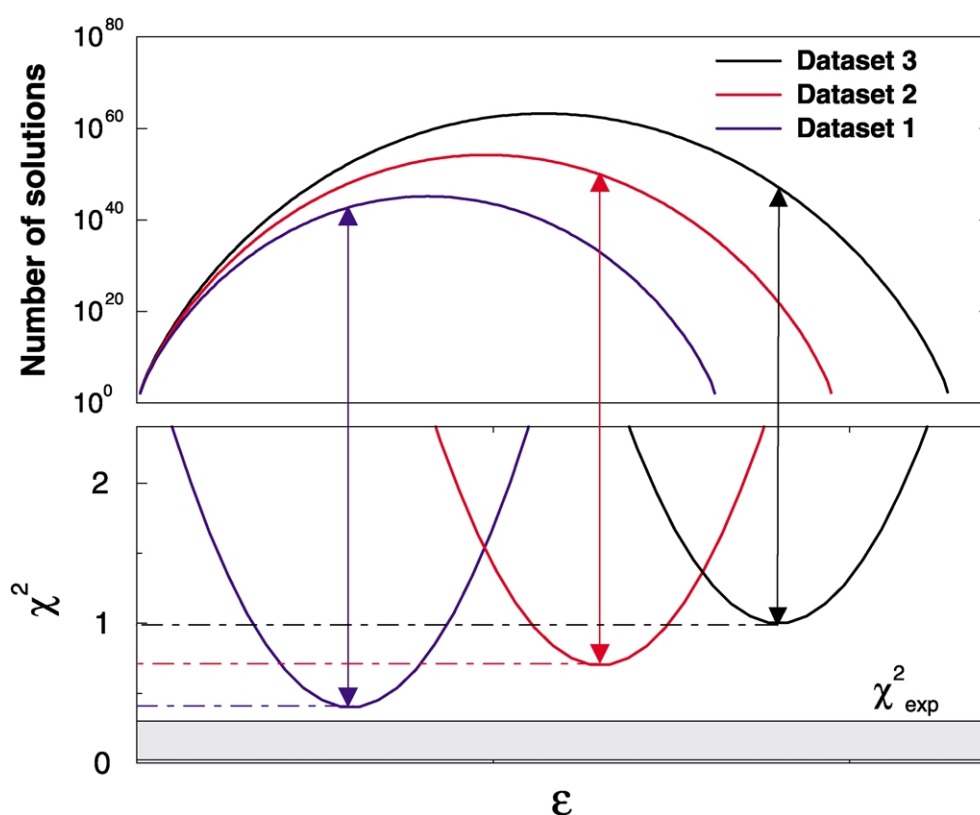

Figure 7. A cartoon representation of the balance between uniqueness and accuracy of solutions obtained. The parameter ε determines the number of singular values used in SVD and, consequently, the rank of the contact matrix A (see Methods). For any set of data, the number of solutions (top panel) is minimal when ε is close to 0 and when ε is close to 1. However, the corresponding χ^2 values, a measure of the accuracy of prediction (bottom panel), is much greater than χ^2_{min}. The χ^2 values reach a minimum in the range $0 < \varepsilon < 1$. We obtain multiple solutions at values of ε corresponding to χ^2_{min}. In all cases, the obtained χ^2_{min} is greater than the error in experimental measurements χ^2_{exp}. Thus, we argue that a solution that is both unique and accurate ($\chi^2_{min} \leq \chi^2_{exp}$) cannot be obtained.

For these proteins, the assumption of attributing the $\Delta\Delta G$ exclusively to the change in contact energies from the folded state breaks down. A quantitative understanding of the energetics of the unfolded states may therefore be necessary for understanding protein stability.

Conclusions

We develop a new methodology for deriving contact interaction parameters to predict stability changes in proteins upon mutation ($\Delta\Delta G$). Using experimental $\Delta\Delta G$ measurements, we derive contact interaction parameters for energy functions that include two-body, three-body and solvation contributions to protein stability, by solving a set of linear equations. Each linear equation relates the changes in the amino acid contact composition of the protein upon mutation to the experimentally measured $\Delta\Delta G$ value. We obtain a reasonable correlation between experimental and predicted $\Delta\Delta G$ with a simple two-body contact approximation and find that the inclusion of solvation and multibody interactions does not lead to a significant improvement in predictive capability of our models. The differences in the predictive capabilities are related to the frequency of occurrence of contact parameters in the set of linear equations being solved. Therefore, rationally designed mutant libraries, intended to maximally cover parameter space, will aid the estimation of contact parameters. However, we argue that even for well-represented datasets, the $\Delta\Delta G$ predictions may not be adequately accurate. For all datasets in our study, corresponding to $\Delta\Delta G$ measurements for individual proteins (eglin c, staphylococcal nuclease) and for a set of 11 proteins, we find that the uncertainties in predictions made by using the contact approximation are always greater than the corresponding experimental errors. Furthermore, we compare interaction parameters derived from the three datasets and find negligible correlation between them. We show that the number of optimal solutions that satisfy all three datasets is always greater than one. Thus, we argue that it may not be possible to determine universal, fully transferable contact interaction parameters that will be able to predict $\Delta\Delta G$ accurately and simultaneously for a set of proteins. Our results suggest that a more atomistic form of potential and/or inclusion of the unfolded state into contact models may be necessary for both an accurate estimation of protein stability and achieving transferability of derived potentials across a set of proteins.

Methods

Potential models

We use a hierarchical approach to develop a simplified model of amino acid interactions. We approximate the potential energy of amino acid interactions in a protein by the sum of terms:

$$H = H_2 + H_3 + H_s\ldots \qquad (2)$$

where H_2 and H_3 are the contributions from two-body and three-body interactions, respectively, H_s is the contribution to the protein potential energy from the solvation of amino acid residues.

Two-body Hamiltonian

We use an interaction model for two-body interaction $H = H_2$, assuming effective interactions between pairs of C^β atoms. The two-body Hamiltonian $H_2(\Gamma,S)$ for a protein of conformation Γ with a given sequence of amino acids S is the sum of the pairwise interaction potentials $u_2(\sigma_a,\sigma_b)$, between amino acids of types σ_a and σ_b:

$$H_2(\Gamma,S) = \sum_{\sigma_a,\sigma_b}^{M} \sum_{i\neq j=1}^{N} u_2(\sigma_a,\sigma_b)s_i(\sigma_a)s_j(\sigma_b)\Delta_{ij} \qquad (3)$$

where $s_i(\sigma_a) \equiv \delta(\sigma_a - \sigma_i)$ and σ_i is the amino acid type at position i along the sequence $S = \{\sigma_1,\sigma_2,\ldots,\sigma_N\}$; $\delta(x)$ is 1 if $x = 0$ or 0 otherwise; $\|\Delta_{ij}\|$ is the contact matrix whose elements Δ_{ij} are either 1 or 0 depending on whether amino acid residues at the positions i and j along the sequence are in contact. The contact between amino acid residues i and j is defined to exist if their corresponding C^β atoms (C^α in case of Gly) are separated by a distance of 7.5 Å.[38] M is the total number of distinct amino acid types and N is the length (number of amino acid residues) of the protein. For the two-body term, we account for all $M = 20$ types of amino acids.

Solvation Hamiltonian

Proteins are affected strongly by the solvent, so it is important to take into account the contribution of the solvation energy term, H_s to the total amino acid interactions.[53] We use an approximate method for the determination of the solvent contribution originally proposed by Eisenberg,[56] which assumes that the effect of solvent is proportional to the solvent-accessible surface area buried upon folding. We approximate the solvent contribution as:

$$H_s = \sum_{\sigma_a}^{M} \sum_{i}^{N} u_s(\sigma_a)s_i(\sigma_a)\Sigma_i, \qquad (4)$$

where $u_s(\sigma_a)$ is the solvent energy term for the amino acid residue of type σ_a, and Σ_i is solvent-accessible surface area for the ith residue. To compute H_s we calculate the solvent-accessible surface area for each amino acid residue by estimating the surface area swept by the center of a solvent probe molecule (modeled as a rigid sphere of finite radius) when it rolls about the van der Waals surface of the protein. The van der Waals surface is the external surface of the atoms, each represented by a spherical ball of its van der Waals radius. The methodology to perform solvent-accessible surface area computations is well established.[56,68–70] We use the freely available program GETAREA1.1 to estimate the solvent-accessible surface area of a residue.[68]

Three-body Hamiltonian

It has been shown that non-additive protein interactions

play an important role in protein folding.[71,54] We account for the non-additive interactions by adding terms corresponding to three-body and higher-order terms. Using the same notation as in equation (3), we approximate the three-body interaction of amino acid residues as:

$$H_3 = \sum_{\sigma_a,\sigma_b,\sigma_c} \sum_{i \neq j \neq k=1}^{N} u_3(\sigma_a,\sigma_b,\sigma_c) s_i(\sigma_a) s_j(\sigma_b) s_k(\sigma_c) \Delta_{ij}\Delta_{jk}\Delta_{ki},$$

(5)

where $u_3(\sigma_a,\sigma_b,\sigma_c)$ are the interaction parameters between residues of types σ_a, σ_b, and σ_c. Equation (5) requires the formation of a triad of contacts simultaneously between amino acid residues at positions i,j, and k, so that H_3 contributes to H only when there are triads of residues that are in contact with each other. The H_3 term has, therefore, an additional contribution favoring or disfavoring triad formations. The H_3 and higher-order non-additive terms increase the number of interaction parameters drastically: for $M = 20$, the number of interaction parameters in H_2 (equation (3)) is 210, while the number of parameters in H_3 (equation (5)) is 1540. We also use a reduced alphabet of amino acids, $M = 6$ in H_3.[57] In the reduced alphabet, we group amino acids into aliphatic (A, V, L, I, M and C), aromatic (F, W, Y and H), polar (S, T, N and Q), positively charged (K and R), negatively charged (D and E) and special (reflecting their special conformational properties) (G and P).[57,72] We have a total of 56 reduced-alphabet, three-body interaction parameters.

Deriving the interaction parameters

To find the interaction parameters $((u_2(\sigma_a,\sigma_b), u_3(\sigma_a,\sigma_b,\sigma_c), u_s(\sigma_a),...)$ in equations (3)–(5) we use experimentally determined changes in the thermodynamic stabilities of proteins upon mutations ($\Delta\Delta G$) from a large number of mutant proteins. Our derivation is based on the approximation that the changes in entropies between mutant and wild-type proteins arise predominantly due to the differences in the solvation free energies of the mutant and wild-type amino acid residues, so that the structures of both mutant and wild-type proteins remain close to each other (within 2 Å RMSD), enough to assume that the change in conformational entropy is zero (we note that the solvation free energy difference can be accounted for by the term H_s (equation (4)). Thus, the difference between stabilities of mutant and wild-type proteins is due to the change in contact energies arising due to mutation. As contact energies are effective free energies, from the quasi-chemical approximation, we can equate the experimentally determined changes in thermodynamic stabilities $\Delta\Delta G$ to $\Delta\Delta H$ values from our model:

$$\Delta\Delta G \approx \Delta\Delta H = \Delta H_{\text{mut}} - \Delta H_{\text{wt}} = H_{\text{mut}} - H_{\text{mut}}^{\text{u}} - H_{\text{wt}} + H_{\text{wt}}^{\text{u}}$$

where H_{mut}, $H_{\text{mut}}^{\text{u}}$, H_{wt}, and H_{wt}^{u} are the energies of mutant native and unfolded, and wild-type native and unfolded states, respectively.

We assume that the number of contacts in the unfolded state is much smaller than those in the native state. We assume also that the structure of the unfolded state is not affected by mutation significantly. Therefore, the difference in energies of the unfolded states can be approximated by the changes in the solvation free energies of the wild-type and mutant amino acid residues, accounted for by equation (4), and we neglect the differ-

ence $H_{\text{mut}}^{\text{u}} - H_{\text{wt}}^{\text{u}}$ to obtain:

$$\Delta\Delta G = H_{\text{mut}} - H_{\text{wt}}$$

(6)

By substituting equations (3)–(5) to equation (6), for each $\Delta\Delta G$ measurement, we obtain a linear equation relating experimentally determined free energy change and contact energy parameters of our model with a given alphabet size (e.g. 210 two-body, 1540 three-body and 20 solvation interaction parameters for a 20-letter amino acid alphabet). For all the measurements in a particular dataset, we obtain a set of linear equations relating the free energy changes and the contact interaction parameters. We simultaneously solve the set of linear equations using SVD corresponding to a large number of mutants to determine these interaction parameters.

Singular value decomposition

Since the number of $\Delta\Delta G$ measurements may not be exactly equal to the total number of variables (the interaction parameters $u_2(\sigma_a,\sigma_b),u_3(\sigma_a,\sigma_b,\sigma_c),u_s(\sigma_a),...$ in equations (3)–(5)), it is expected that exact solutions are not achievable. Instead, we find the linear least-squares solution of equations (6). The linear least-squares problem for a set of linear equations, $A\mathbf{x} = \mathbf{b}$ where A is a known $m \times n$ matrix, $\mathbf{b}$ is a known m-dimensional vector and $\mathbf{x}$ is an n-dimensional solution vector, is defined as finding a solution vector $\mathbf{x}$ that minimizes the L_2 norm of the residual vector $\|A\mathbf{x} - \mathbf{b}\|_2$. If $m \geq n$ and rank$(A) = n$, we find a unique least-squares solution. If $m < n$, there is an infinite number of least-squares solutions, so we seek the minimum norm least-squares solution that minimizes both $\|\mathbf{x}\|_2$ and $\|A\mathbf{x} - \mathbf{b}\|_2$.

To find the solution for both the above cases, we implement the SVD algorithm.[73] The SVD of an $m \times n$ matrix A is given by $A = UWV^{\text{T}}$, where U and V are orthogonal matrices and W is a diagonal matrix whose elements w_i, known as singular values of A, satisfy $w_1 \geq w_2 \geq ... \geq w_{\text{min}} \geq 0$. The SVD algorithm ensures a unique decomposition of this matrix. In our case, A corresponds to the matrix of contact interactions, $\mathbf{x}$ is the solution vector of contact energies and $\mathbf{b}$ is the experimentally determined $\Delta\Delta G$ values. The dimension of $\mathbf{x}$ and $\mathbf{b}$ are the number of parameters (N_{p}) to be determined and the number of experimental measurements (N_{e}), respectively. We identify an approximate solution that satisfies equations (6) for all the mutants simultaneously so that:

$$\chi^2 \equiv \sum_{\alpha=1}^{N_{\text{e}}} \|\Delta\Delta G - \Delta\Delta H\|_{\alpha}^2$$

(7)

is minimized ($\chi_{\text{min}}^2 = \min \chi^2$). In equation (7) α enumerates N_{e} experimental measurements.

In the SVD algorithm, we decompose the matrix A into its singular values, $\{w_i\}$. If the matrix is singular we find that some w_i are zero. If the matrix A is not singular, but ill-conditioned (i.e. some w_i values cannot be determined accurately), then the solution vector may have a large error-prone component that yields a non-least-squares solution ($\chi^2 \neq \chi_{\text{min}}^2$). Therefore, in order to find the least-square solution for this case we define the relative singular value, $\varepsilon_i = w_i/w_{\text{max}}$ (w_{max} is the maximum singular value in magnitude). If ε_i is smaller than some threshold value, ε, we neglect the corresponding w_i for finding the solution. Thus, ε determines the number of non-zero singular values in (and consequently the rank of) the design matrix A and hence the solution vector $\mathbf{x}$. In order to obtain the most accurate solution, we choose

the value of ε such that the value of χ^2 is the minimal (χ^2_{min}). If multiple solutions correspond to a value of χ^2_{min}, we choose the minimal norm (minimizing the length of $\bar{x}$) solution.

Tests of reliability of models

Resubstitution test

The resubstitution test is an examination for the self-consistency of a prediction algorithm. In the resubstitution test $\Delta\Delta G$ for each mutant are predicted for the dataset (testing set) from which the model parameters are derived (training set). The resubstitution test sets an upper bound on the prediction ability of the derived potential when the training set and the testing set are not identical. The performance of a model in the resubstitution test is an optimistic estimation.[74–76] Thus, the resubstitution test is necessary but not sufficient for evaluating a prediction method. As a complement, a cross-validation test is needed because it would reflect the effectiveness of a prediction algorithm.

Jack-knife test

We perform a jack-knife test, known also as a leave-one-out test,[77] to cross-validate the statistical prediction of our interaction model. The jack-knife provides an objective assessment of the predictive capabilities of a model.[78] In the jack-knife test, the experimental data of N points is divided into two groups: the training set consisting of $N - 1$ data points and the testing set consisting of one data point. We derive the interaction model parameters using the training set data and predict the value of $\Delta\Delta G$ for the testing set. Then another data point is selected from the experimental data to represent the testing set, while the remaining $N - 1$ data points represent the training set. We repeat this procedure for all N data points. The training and testing thus covers all of the experimental data points.

Acknowledgements

References

1. Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science*, **181**, 223–230.
2. Fersht, A. R. & Shakhnovich, E. I. (1998). Protein folding: think globally (inter)act locally. *Curr. Biol.* **8**, R478–R479.
3. Levitt, M., Gerstein, M., Huang, E., Subbiah, S. & Tsai, J. (1997). Protein folding: the endgame. *Annu. Rev. Biochem.* **66**, 549–579.
4. Onuchic, J. N., LutheySchulten, Z. & Wolynes, P. G. (1997). Theory of protein folding: the energy landscape perspective. *Annu. Rev. Phys. Chem.* **48**, 545–600.
5. Onuchic, J. N., Nymeyer, H., Garcia, A. E., Chahine, J. & Socci, N. D. (2000). The energy landscape theory of protein folding: insights into folding mechanisms and scenarios. *Advan. Protein Chem.* **53**, 87–152.
6. Pande, V. S., Grosberg, A. Y. & Tanaka, T. (2000). Heteropolymer freezing and design: towards physical models of protein folding. *Rev. Modern Phys.* **72**, 259–314.
7. Plaxco, K. W., Riddle, D. S., Grantcharova, V. & Baker, D. (1998). Simplified proteins: minimalist solutions to the protein folding problem. *Curr. Opin. Struct. Biol.* **8**, 80–85.
8. Shakhnovich, E. I. (1997). Theoretical studies of protein-folding thermodynamics and kinetics. *Curr. Opin. Struct. Biol.* **7**, 29–40.
9. Dahiyat, B. I. & Mayo, S. L. (1997). *De novo* protein design: fully automated sequence selection. *Science*, **278**, 82–87.
10. Deutsch, J. M. & Kurosky, T. (1996). New algorithm for protein design. *Phys. Rev. Letters*, **76**, 323–326.
11. Hellinga, H. W. (1997). Rational protein design: combining theory and experiment. *Proc. Natl Acad. Sci. USA*, **94**, 10015–10017.
12. Hinds, D. A. & Levitt, M. (1996). From structure to sequence and back again. *J. Mol. Biol.* **258**, 201–209.
13. Jaramillo, A., Wernisch, L., Hery, S. & Wodak, S. J. (2001). Automatic procedures for protein design. *Comb. Chem. High Throughput Screen.* **4**, 643–659.
14. Jin, W., Kambara, O., Sasakawa, H., Tamura, A. & Takada, S. (2003). *De novo* design of foldable proteins with smooth folding funnel. Automated negative design and experimental verification. *Structure (Camb)*, **11**, 581–590.
15. Kamtekar, S., Schiffer, J. M., Xiong, H. Y., Babik, J. M. & Hecht, M. H. (1993). Protein design by binary patterning of polar and nonpolar amino-acids. *Science*, **262**, 1680–1685.
16. Kraemer-Pecore, C. M., Wollacott, A. M. & Desjarlais, J. R. (2001). Computational protein design. *Curr. Opin. Chem. Biol.* **5**, 690–695.
17. Kuhlman, B. & Baker, D. (2000). Native protein sequences are close to optimal for their structures. *Proc. Natl Acad. Sci. USA*, **97**, 10383–10388.
18. Seno, F., Vendruscolo, M., Maritan, A. & Banavar, J. R. (1996). Optimal protein design procedure. *Phys. Rev. Letters*, **77**, 1901–1904.
19. Shakhnovich, E. I. (1998). Protein design: a perspective from simple tractable models. *Fold. Des.* **3**, R45–R58.
20. Yue, K. & Dill, K. A. (1992). Inverse protein folding problem—designing polymer sequences. *Proc. Natl Acad. Sci. USA*, **89**, 4163–4167.
21. Snow, C. D., Nguyen, H., Pande, V. S. & Gruebele, M. (2002). Absolute comparison of simulated and experimental protein-folding dynamics. *Nature*, **420**, 102–106.
22. Zagrovic, B., Snow, C. D., Shirts, M. R. & Pande, V. S. (2002). Simulation of folding of a small alpha-helical protein in atomistic detail using worldwide-distributed computing. *J. Mol. Biol.* **323**, 927–937.
23. Simmerling, C., Strockbine, B. & Roitberg, A. E. (2002). All-atom structure prediction and folding simulations of a stable protein. *J. Am. Chem. Soc.* **124**, 11258–11259.
24. Pitera, J. W. & Swope, W. (2003). Understanding folding and design: replica-exchange simulations of "Trp-cage" fly miniproteins. *Proc. Natl Acad. Sci. USA*, **100**, 7587–7592.
25. Skolnick, J., Kolinski, A., Sikorski, A. & Yaris, R. (1989). Dynamic Monte–Carlo simulations of

globular protein folding. *Abstr. Pap. Am. Chem. Soc.* **197**, 110-HYS.

26. Klimov, D. K. & Thirumalai, D. (1996). Factors governing the foldability of proteins. *Proteins: Struct. Funct. Genet.* **26**, 411–441.

27. Kussell, E., Shimada, J. & Shakhnovich, E. I. (2003). Side-chain dynamics and protein folding. *Proteins: Struct. Funct. Genet.* **52**, 303–321.

28. Dokholyan, N. V., Buldyrev, S. V., Stanley, H. E. & Shakhnovich, E. I. (1998). Discrete molecular dynamics studies of the folding of a protein-like model. *Fold. Des.* **3**, 577–587.

29. Voegler, A. C. & Hall, C. K. (1999). Discontinuous molecular dynamics simulations of simple protein systems. *Biophys. J.* **76**, A153.

30. Zhou, Y. Q., Hall, C. K. & Karplus, M. (1996). First-order disorder-to-order transition in an isolated homopolymer model. *Phys. Rev. Letters*, **77**, 2822–2825.

31. Zhou, Y. Q. & Karplus, M. (1999). Folding of a model three-helix bundle protein: a thermodynamic and kinetic analysis. *J. Mol. Biol.* **293**, 917–951.

32. Clementi, C., Maritan, A. & Banavar, J. R. (1998). Folding, design, and determination of interaction potentials using off-lattice dynamics of model heteropolymers. *Phys. Rev. Letters*, **81**, 3287–3290.

33. Dima, R. I., Settanni, G., Micheletti, C., Banavar, J. R. & Maritan, A. (2000). Extraction of interaction potentials between amino acids from native protein structures. *J. Chem. Phys.* **112**, 9151–9166.

34. Go, N. (1983). Theoretical studies of protein folding. *Annu. Rev. Biophys. Bioeng.* **12**, 183–210.

35. Maiorov, V. N. & Crippen, G. M. (1992). Contact potential that recognizes the correct folding of globular-proteins. *J. Mol. Biol.* **227**, 876–888.

36. Micheletti, C., Seno, F., Banavar, J. R. & Maritan, A. (2001). Learning effective amino acid interactions through iterative stochastic techniques. *Proteins: Struct. Funct. Genet.* **42**, 422–431.

37. Miyazawa, S. & Jernigan, R. L. (1985). Estimation of effective interresidue contact energies from protein crystal-structures-quasi-chemical approximation. *Macromolecules*, **18**, 534–552.

38. Miyazawa, S. & Jernigan, R. L. (1996). Residue–residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.* **256**, 623–644.

39. Ding, F., Dokholyan, N. V., Buldyrev, S. V., Stanley, H. E. & Shakhnovich, E. I. (2002). Direct molecular dynamics observation of protein folding transition state ensemble. *Biophys. J.* **83**, 3525–3532.

40. Vendruscolo, M. & Domany, E. (1998). Pairwise contact potentials are unsuitable for protein folding. *J. Chem. Phys.* **109**, 11101–11108.

41. Vendruscolo, M., Najmanovich, R. & Domany, E. (1999). Protein folding in contact map space. *Phys. Rev. Letters*, **82**, 656–659.

42. Domingues, H., Peters, J., Schneider, K. H., Apeler, H., Sebald, W., Oschkinat, H. & Serrano, L. (2000). Improving the refolding yield of interleukin-4 through the optimization of local interactions. *J. Biotechnol.* **84**, 217–230.

43. 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.

44. Lazaridis, T. & Karplus, M. (2000). Effective energy functions for protein structure prediction. *Curr. Opin. Struct. Biol.* **10**, 139–145.

45. Takano, K., Ota, M., Ogasahara, K., Yamagata, Y., Nishikawa, K. & Yutani, K. (1999). Experimental verification of the stability profile of mutant protein (SPMP) data using mutant human lysozymes. *Protein Eng.* **12**, 663–672.

46. Edgell, M. H., Sims, D. A., Pielak, G. J. & Yi, F. (2003). High-precision, high-throughput stability determinations facilitated by robotics and a semiautomated titrating fluorometer. *Biochemistry*, **42**, 7587–7593.

47. Gromiha, M. M., An, J., Kono, H., Oobatake, M., Uedaira, H. & Sarai, A. (1999). ProTherm: thermodynamic database for proteins and mutants. *Nucl. Acids Res.* **27**, 286–288.

48. Zhou, H. Y. & Zhou, Y. Q. (2002). Stability scale and atomic solvation parameters extracted from 1023 mutation experiments. *Proteins: Struct. Funct. Genet.* **49**, 483–492.

49. Chou, K. C., Liu, W. M., Maggiora, G. M. & Zhang, C. T. (1998). Prediction and classification of domain structural classes. *Proteins: Struct. Funct. Genet.* **31**, 97–103.

50. Bhat, M. G., Ganley, L. M., Ledman, D. W., Goodman, M. A. & Fox, R. O. (1997). Stability studies of amino acid substitutions at tyrosine 27 of the staphylococcal nuclease beta-barrel. *Biochemistry*, **36**, 12167–12174.

51. 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.

52. Stites, W. E., Gittis, A. G., Lattman, E. E. & Shortle, D. (1991). In a staphylococcal nuclease mutant the side-chain of a lysine replacing valine-66 is fully buried in the hydrophobic core. *J. Mol. Biol.* **221**, 7–14.

53. Dill, K. A. (1990). Dominant forces in protein folding. *Biochemistry*, **29**, 7133–7155.

54. Chen, J. M. & Stites, W. E. (2001). Higher-order packing interactions in triple and quadruple mutants of staphylococcal nuclease. *Biochemistry*, **40**, 14012–14019.

55. Carter, C. W., Jr, LeFebvre, B. C., Cammer, S. A., Tropsha, A. & Edgell, M. H. (2001). Four-body potentials reveal protein-specific correlations to stability changes caused by hydrophobic core mutations. *J. Mol. Biol.* **311**, 625–638.

56. Eisenberg, D., Wilcox, W. & Mclachlan, A. D. (1986). Hydrophobicity and amphiphilicity in protein-structure. *J. Cell. Biochem.* **31**, 11–17.

57. Dokholyan, N. V. (2004). What is the protein design alphabet? *Proteins: Struct. Funct. Bioinformat.* **54**, in press.

58. Buchler, N. E. G. & Goldstein, R. A. (1999). Effect of alphabet size and foldability requirements on protein structure designability. *Proteins: Struct. Funct. Genet.* **34**, 113–124.

59. Fan, K. & Wang, W. (2003). What is the minimum number of letters required to fold a protein? *J. Mol. Biol.* **328**, 921–926.

60. Yi, F., Sims, D. A., Pielak, G. J. & Edgell, M. H. (2003). Testing hypotheses about determinants of protein structure with high-precision, high-throughput stability measurements and statistical modeling. *Biochemistry*, **42**, 7594–7603.

61. Holder, J. B., Bennett, A. F., Chen, J. M., 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.

62. Meeker, A. K., GarciaMoreno, B. & Shortle, D. (1996).

Contributions of the ionizable amino acids to the stability of staphylococcal nuclease. *Biochemistry,* **35**, 6443–6449.

63. Sanz, J. M. & Fersht, A. R. (1993). Rationally designing the accumulation of a folding intermediate of barnase by protein engineering. *Biochemistry,* **32**, 13584–13592.

64. Shortle, D. & Ackerman, M. S. (2001). Persistence of native-like topology in a denatured protein in 8 M urea. *Science,* **293**, 487–489.

65. Shortle, D. & Meeker, A. K. (1989). Residual structure in large fragments of staphylococcal nuclease—effects of amino-acid substitutions. *Biochemistry,* **28**, 936–944.

66. Shortle, D. (1993). Denatured states of proteins and their roles in folding and stability. *Curr. Opin. Struct. Biol.* **3**, 66–74.

67. Shortle, D. (1996). The denatured state (the other half of the folding equation) and its role in protein stability. *FASEB J.* **10**, 27–34.

68. Fraczkiewicz, R. & Braun, W. (1998). Exact and efficient analytical calculation of the accessible surface areas and their gradients for macromolecules. *J. Comput. Chem.* **19**, 319–333.

69. Ooi, T., Oobatake, M., Nemethy, G. & Scheraga, H. A. (1987). Accessible surface-areas as a measure of the thermodynamic parameters of hydration of peptides. *Proc. Natl Acad. Sci. USA,* **84**, 3086–3090.

70. Vorobjev, Y. N. & Hermans, J. (1997). SIMS: computation of a smooth invariant molecular surface. *Biophys. J.* **73**, 722–732.

71. Chen, J. M. & Stites, W. E. (2000). Additivity of double mutant stability effects in the core of staphylococcal nuclease. *Biophys. J.* **78**, 2487Pos.

72. Dokholyan, N. V. & Shakhnovich, E. I. (2001). Understanding hierarchical protein evolution from first principles. *J. Mol. Biol.* **312**, 289–307.

73. Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. (2002). *Numerical Recipes in C*, 2nd edit., Chap. 2, Cambridge University Press, Cambridge.

74. Cai, Y. D. (2001). Is it a paradox or misinterpretation? *Proteins: Struct. Funct. Genet.* **43**, 336–338.

75. Chou, K. C. (1995). A novel-approach to predicting protein structural classes in a (20-1)-D amino-acid-composition space. *Proteins: Struct. Funct. Genet.* **21**, 319–344.

76. Zhou, G. P. & Assa-Munt, N. (2001). Some insights into protein structural class prediction. *Proteins: Struct. Funct. Genet.* **44**, 57–59.

77. Klein, P. & Delisi, C. (1986). Prediction of protein structural class from the amino-acid-sequence. *Biopolymers,* **25**, 1659–1672.

78. Chou, K. C. & Zhang, C. T. (1995). Prediction of protein structural classes. *Crit. Rev. Biochem. Mol. Biol.* **30**, 275–349.

79. Sarai, A., Gromiha, M. M., An, J., Prabakaran, P., Selvaraj, S., Kono, H. *et al.* (2002). Thermodynamic databases for proteins and protein–nucleic acid interactions. *Biopolymers,* **61**, 121–126.

Edited by G. von Heijne