doi:10.1016/S0022-2836(02)00442-4 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2002) 320, 369–387

JMB

Predicting Changes in the Stability of Proteins and Protein Complexes: A Study of More Than 1000 Mutations

Raphael Guerois[1,2]*, Jens Erik Nielsen[1,3] and Luis Serrano[1]

[1]*EMBL, Meyerhofstrasse 1 69117 Heidelberg, Germany*

[2]*SBFM-DBJC, Centre d'Etude of Saclay, 91191 Gif sur Yvette Cedex, France*

[3]*Howard Hughes Medical Institute and Department of Chemistry and Biochemistry MC 0365, University of California, San Diego, La Jolla CA 92093, USA*

**Corresponding author*

We have developed a computer algorithm, FOLDEF (for FOLD-X energy function), to provide a fast and quantitative estimation of the importance of the interactions contributing to the stability of proteins and protein complexes. The predictive power of FOLDEF was tested on a very large set of point mutants (1088 mutants) spanning most of the structural environments found in proteins. FOLDEF uses a full atomic description of the structure of the proteins. The different energy terms taken into account in FOLDEF have been weighted using empirical data obtained from protein engineering experiments. First, we considered a training database of 339 mutants in nine different proteins and optimised the set of parameters and weighting factors that best accounted for the changes in stability of the mutants. The predictive power of the method was then tested using a blind test mutant database of 667 mutants, as well as a database of 82 protein–protein complex mutants. The global correlation obtained for 95 % of the entire mutant database (1030 mutants) is 0.83 with a standard deviation of 0.81 kcal mol^{-1} and a slope of 0.76. The present energy function uses a minimum of computational resources and can therefore easily be used in protein design algorithms, and in the field of protein structure and folding pathways prediction where one requires a fast and accurate energy function. FOLDEF is available *via* a web-interface at http://fold-x.embl-heidelberg.de

Keywords: protein stability; mutation; thermodynamic; prediction; protein complex

Introduction

The translation of structural data into energetic parameters is one of the long-term goals of protein structure analysis. Moreover, there is a need for accurate and fast algorithms for protein energy calculations, in particular for the development of algorithms with complex search procedures and numerous combinatorial calculations. Typical examples of such algorithms are protein design and protein docking algorithms[1,2] and protein structure prediction methods.[3] Recently, the possibility of predicting protein folding pathways from their folded structure prompted an interest in obtaining fast and reliable energy calculations from a static protein structure.[4–7]

The development of a fast and reliable protein force-field is a complex task, given the delicate balance between the different energy terms that contribute to protein stability.[8,9] Many different force-fields have been constructed for predicting protein stability changes. These range from force-fields based on pure statistical analysis of structural sequence preferences,[10–14] and force-fields based on multiple sequence alignments,[15–17] to detailed molecular dynamics force-fields.[18,19]

These force-fields can be divided into three major categories: (i) those using a physical effective energy function (PEEF); (ii) those based on statistical potentials for which energies are derived from the frequency of residue or atom contacts in the protein database (SEEF) as reviewed by Lazaridis & Karplus;[3] and (iii) those using empirical data obtained form experiments ran on proteins (EEEF).

Abbreviations used: ΔG, free energy difference between the folded and unfolded protein; ΔG_{KD}, binding free energy difference between the protein complex and the unbound states; $\Delta\Delta G$, free energy difference between the wild-type (WT) and the mutant protein.

E-mail address of the corresponding author: guerois@cea.fr

0022-2836/02/$ - see front matter

The main drawbacks of the PEEF potentials are that they are computationally very expensive and they can therefore be used only on small sets of protein mutants. The computation time can be reduced somewhat by using implicit terms for solvation energies and side-chain entropies, but the time required to get a reliable estimate of a free energy difference between a wild-type and mutant protein is still significant.[20]

The power of SEEFs is that they contain terms that account for complex effects that are difficult to describe separately, and they contain empirical approximations for the denatured state. A drawback of this approach is that once an SEEF potential has been constructed, improvements cannot be added easily without introducing overlaps in the underlying energies.

EEEF approaches combine a physical description of the interactions with lessons learned from experiments. Good examples of such algorithms are the helix/coil transition algorithm AGADIR[21,22] or the SPMP method.[23] The AGADIR algorithm is accurate at predicting the helical content of peptides in solution and has been used to design mutations that increase the thermostability of a protein through local interactions.[24–26] A limitation of this algorithm is that it can be applied only to α-helices and cannot take tertiary interactions into account.

Here, we have developed an energy function based on the EEEF approach using a strategy similar to that used for the development of AGADIR. We have taken advantage of the large amount of experimental work that has been devoted to understanding protein energetics. In particular, we have relied on the body of data that probed, through single and multiple-residue mutation analysis, the roles of particular interactions that contribute to protein stability.[27,28] We followed a two-step procedure. First, we considered a training database of 339 mutants in nine different proteins and optimised the set of parameters and weighting factors that best accounted for the changes in stability of the mutants. The predictive power of the method was then tested using a blind test mutant database of 667 mutants, as well as a database of 82 protein–protein complex mutants.

Considering the training and the blind test database together, the algorithm was tested over 1088 mutants. In this entire database, most of the important interactions that govern protein stability are represented in the protein mutant database. All types of secondary structures are represented substantially (turn, 17%; alpha and 3_{10}-helix, 30%; β-sheet, 32%; coil, 21%). There is a similar number of mutations that involve only hydrophobic residues and mutations that involve deletions or substitution of polar atoms (47% and 53%, respectively). Finally, the percentages of mutated residues having a solvent-accessibility higher or lower than 30% are similar, 45% and 55% of the mutant database, respectively. The global correlation obtained for 95% of the entire mutant database is 0.83 with a standard deviation of 0.81 kcal mol^{-1} (1 cal = 4.184 J) and a slope of 0.76. The present energy function FOLDEF (FOLD-X energy function, in the following) uses a minimum of computational resources and can therefore be used easily in protein design algorithms where one requires a fast and accurate energy function.

Results

Energy terms in the FOLD-X energy function

The FOLD-X energy function (FOLDEF) includes terms that have been found to be important for protein stability. The free energy of unfolding (ΔG) of a target protein is calculated using equation (1):

$$\Delta G = W_{vdw}\Delta G_{vdw} + W_{solvH}\Delta G_{solvH} + W_{solvP}\Delta G_{solvP} + \Delta G_{wb} + \Delta G_{hbond} + \Delta G_{el} + W_{mc}T\Delta S_{mc} + W_{sc}T\Delta S_{sc} \quad (1)$$

where ΔG_{vdw} is the sum of the van der Waals contributions of all atoms. ΔG_{solvH} and ΔG_{solvP} is the difference in solvation energy for apolar and polar groups, respectively, when going from the unfolded to the folded state. ΔG_{hbond} is the free energy difference between the formation of an intra-molecular hydrogen-bond compared to inter-molecular hydrogen-bond formation (with solvent). ΔG_{wb} is the extra stabilising 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.[29–31] ΔG_{el} is the electrostatic contribution of charged groups interactions. ΔS_{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.[32] Finally, ΔS_{sc} is the entropic cost of fixing a side-chain in a particular conformation.[33] The energy values of ΔG_{vdw}, ΔG_{solvH}, ΔG_{solvP} and ΔG_{hbond} attributed to each atom type have been derived from a set of experimental data, and ΔS_{mc} and ΔS_{sc} have been taken from theoretical estimates (see Materials and Methods for details). The terms W_{vdw}, W_{solvH}, W_{solvP}, W_{mc} and W_{sc} correspond to the weighting factors applied to the raw energy terms. These weights were obtained from an initial fitting procedure over a database consisting of 339 single point mutants (see Materials and Methods).

Effect of solvent exposure determined by the atomic occupancy (Occ)

Many experimental studies show that interactions at the surface of a protein usually contribute less to the stability of a protein than those in the core.[34,35] This can be rationalised as an effect of increased flexibility at the protein surface in an

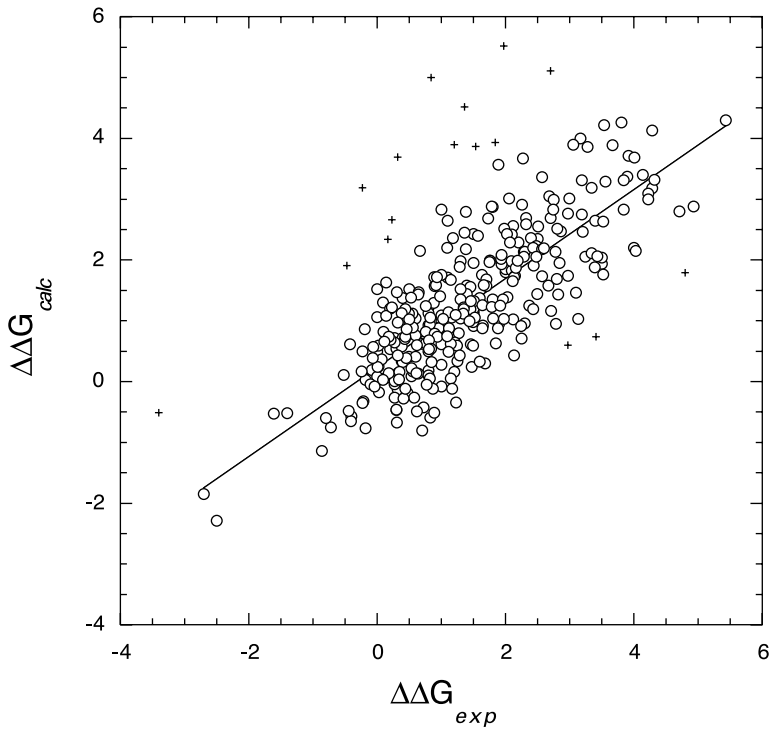

Figure 1. Calculated ΔΔ*G*s compared to the experimental ΔΔ*G*s for the 339 mutants of the training database. The continuous line represents the linear regression obtained with 95% of the training database after the outliers (see Results) were discarded. Its equation is $y = 0.24 + 0.73x$, with a correlation factor of 0.8. The mutants considered as outliers are indicated as crosses and the other mutants are shown as circles.

environment close to that of the unfolded state. Therefore, an important part of the energy calculation is based on the inclusion of solvent effects in an implicit manner, except in the special case of water bridges. To estimate the solvent-accessibility of a given atom, we used the solvent contact model,[36] which considers the volume occupied by protein atoms around the atom, called the occupancy (Occ). The occupancy of a given atom i, Occ(i), is the sum of the fragmental volumes of the atoms surrounding this atom within a threshold distance of 6 Å (see details in Materials and Methods).[36–38] This quantity was preferred to the geometrical surface calculation used in the traditional ASA calculation,[39] since it is much faster to calculate.[37]

In FOLDEF, the atomic free energy of solvation, the van der Waals and the electrostatic interactions together with the entropic terms are scaled with respect to the atomic occupancies (Occ). As a first approximation, we assume that the strength of an interaction (solvation effects, van der Waals or electrostatic) and the entropic cost for fixing the conformation of a residue should vary linearly with the atomic occupancy Occ(i). For each atom i, the unscaled energy terms are multiplied by the scaling factor ($S_{\text{fact}}(i)$) that is calculated from the atomic occupancy Occ(i) as:

$$S_{\text{fact}}(i) = \frac{\text{Occ}(i) - \text{Occ}_{\text{min}}(t_i)}{\text{Occ}_{\text{max}}(t_i) - \text{Occ}_{\text{min}}(t_i)}$$

where $\text{Occ}_{\text{min}}(i)$ and $\text{Occ}_{\text{max}}(i)$ are the minimal and maximal occupancies of an atom of type t_i as estimated by Holm & Sander[37] (see Table 3 in the Supplementary Material).

For the main-chain and side-chain entropy, which are calculated at the residue level and not at the atomic level, we consider the mean value of the occupancies of the atoms that compose the main chain and the side-chain, respectively.

Fitting of the weights in equation (1)

The weights applied to the different energy terms in equation (1) were fitted using the experimental ΔΔ*G* values of an initial mutant database comprising 339 single-point mutants in nine different proteins: barnase,[34] CI-2,[40] spectrin,[41] Src SH3,[42] Sso7d,[4] tenascin,[43] FKBP,[44] Ada2h[45] and CheY.[46] The fitting procedure also involved the estimation of the ΔG_{hbond} values. The problems related to the modelling of the mutated side-chain were avoided by considering only mutations involving the deletion of groups in the side-chain. We assume that these mutations do not affect the conformation of the protein drastically. Based on the same assumption, we considered mutations that involved the substitution of groups, such as E → Q, D → N or T → V and the reverse of these.

The various steps of the fitting procedure are described in detail in Materials and Methods. After several iterations, the set of weights that was found to be optimal was $W_{\text{vdw}} = 0.2$, $W_{\text{solvH}} = 1.4$, $W_{\text{solvP}} = 1.25$, $W_{\text{mc}} = 1.0$ and $W_{\text{sc}} = 0.75$. The optimal value for the formation of a hydrogen bond, ΔG_{hbond}, as found to be -1.3 kcal mol^{-1} if the hydrogen bond was formed between two polar groups and -1.4 kcal mol^{-1} if the hydrogen bond was between a polar and a charged group. The physical interpretation of these weights and of their values is provided in Discussion.

Table 1. Analysis of the outliers in the training database

PDB	Mut	ASA (%Acc)	DSSP	$\Delta\Delta G_{exp}$ (kcal mol^{-1})	$\Delta\Delta G_{calc}$ (kcal mol^{-1})	Possible origin of the discrepancy
1A2P	D54A	22.4	S	2.97	0.6	No interpretation
1A2P	N58D	19.4	C	−0.47	1.91	Type I turn pos. *i*. No interpretation
1BF4	V23A	1.1	S	0.23	2.66	Large hydrophobic buried; possible structural relaxation
1BF4	F32A	0.4	S	2.70	5.11	Large hydrophobic buried; core rearrangement[47]
2CHF	V54T	0.0	S	4.80	1.79	Insertion of a fully buried polar atoms with no H-bond; underestimated
2CHF	D57A	5.9	S	−3.40	−0.51	Residue in the active site involved in strong repulsive interactions
1YPC	D71A	37.3	C	3.41	0.74	Type I turn pos. *i*. No interpretation
1YPC	P52A	57.1	S	0.17	2.34	No interpretation
1FMK	F10A	1.8	S	0.84	5.00	Large hydrophobic buried; possible structural relaxation
1FMK	F26A	5.3	C	1.97	5.52	Large hydrophobic buried; possible structural relaxation
1FMK	I34A	1.2	C	0.32	3.69	Large hydrophobic buried; possible structural relaxation
1FMK	W43A	13.0	S	1.20	3.90	Large hydrophobic buried; possible structural relaxation
1FMK	I56A	0.0	S	1.84	3.93	Large hydrophobic buried; possible structural relaxation
1FMK	P57A	6.1	S	1.36	4.52	Large hydrophobic buried; possible structural relaxation
1FMK	Y60A	24.7	3	−0.23	3.19	Large hydrophobic buried; possible structural relaxation
1FKB	I91A	5.5	C	1.54	3.87	Large hydrophobic buried; possible structural relaxation

Prediction of mutation free energy changes for the training database

In Figure 1 we show the correlation between the predicted and experimental data for the training database. For the entire training database, we obtain the following correlation parameters: slope = 0.67 and $R = 0.7$. The standard deviation, σ, is 0.97 kcal mol^{-1}. There are 16 clear outliers (5% of the database) for which $\Delta\Delta G$ is wrong by more than 2σ (see Table 1). Removing these mutants from the database improves the correlation significantly: slope = 0.73, $R = 0.8$ and standard deviation = 0.75 kcal mol^{-1}. (The characteristics of the 339 mutants of the training database are provided in Table 1 of the Supplementary Material, including the $\Delta\Delta G_{exp}$ and $\Delta\Delta G_{calc}$, the accessibility and the secondary structure of the mutated residue.)

Analysis of the outliers in the training database

In the following, we investigate whether the deviations observed for the 16 outliers between $\Delta\Delta G_{exp}$ and $\Delta\Delta G_{calc}$ most likely originate from a significant change in the mutant structure or from a systematic error of the energy function. For the majority of the outliers, there is a straightforward explanation of the observed discrepancies, related to putative conformational changes upon mutation (Table 1). In ten of the outliers, a large hydrophobic side-chain is deleted, and it seems that FOLDEF overestimates the destabilising effect in these cases (mean error of 2.9 kcal mol^{-1}). However, many other hydrophobic mutants were predicted correctly (154 with $R = 0.77$, $\sigma = 0.78$ kcal mol^{-1} and mean error of 0.02 kcal mol^{-1}), which shows that the errors are not coming from a general overestimation of the hydrophobic effect. Instead, the discrepancies between $\Delta\Delta G_{exp}$ and $\Delta\Delta G_{calc}$ for the ten outliers are most likely due to the relaxation of the structure that reduces the cavity formed by the removal of the side-chain (this is mainly the case for the seven mutants in src SH3 (1FMK)). For instance, it has been shown recently that the single-point mutation F32A in Sso7d,[47] detected as an outlier, involves a hydrophobic core rearrangement.[35,48]

Two outliers are mutants involving polar interactions that are found to be more stable in the experiments than in the calculations (D57A in CheY and N58D in barnase). The residue D57 is located in the active site of the CheY protein flanked by two negatively charged residues (D12, D13). FOLDEF predicts that the D57A mutation is stabilising, although the amplitude of the effect is underestimated. This can be explained easily by the fact that K109, which makes a salt-bridge with D57, can switch and make a salt-bridge with D12 (as seen in another structure of CheY with PDB code 1CHN). Thus the mutation of D57 will not result in the loss of a salt-bridge interaction, and the stabilising effect of the mutation should be larger than calculated. Concerning N58D, we find that the mutated D58 is unable to form a favourable interaction that N58 was making with the carboxyl group of L63 in an exposed turn. Structural rearrangements that we do not model are likely to compensate for the loss of this interaction. Lastly, three outliers involving polar group deletion or substitution (V54T in CheY, D71A in CI-2, D54A in barnase) have large destabilising effects on the protein (over 3 kcal mol^{-1}), and these are underestimated by FOLDEF. In the V54T mutant, a polar atom is introduced into the core of the protein with no hydrogen bond partner. The discrepancy between experimental and calculated values for this mutation may reflect an underestimation of the solvation penalty associated with the burial of polar atoms. However, considering the entire mutant database, there are 14 other Val to Thr mutations with a solvent-accessibility

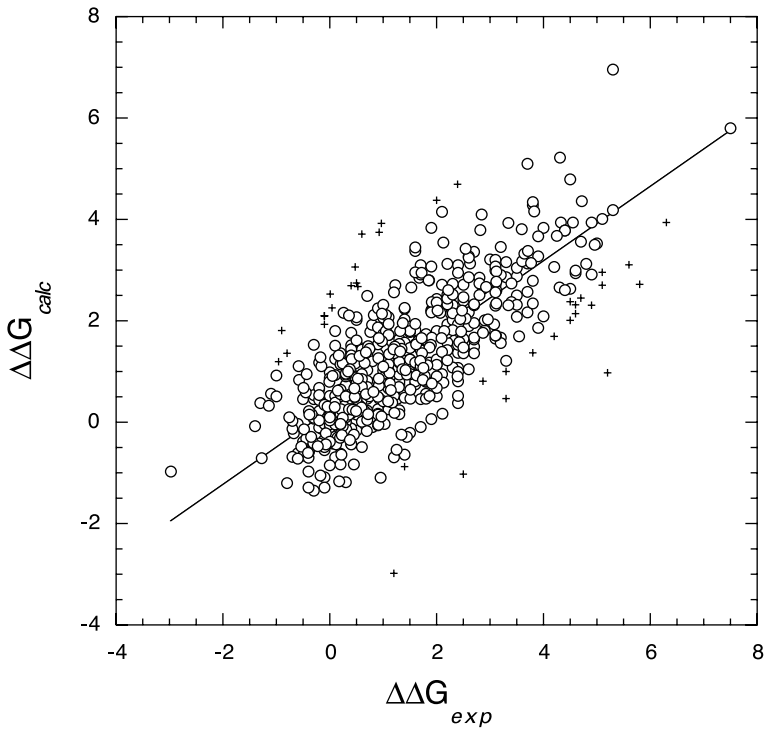

Figure 2. Calculated ΔΔ*G*s compared to the experimental ΔΔ*G*s for the 625 mutants of the blind test database. The continuous line represents the linear regression obtained with 95% of the training database after the outliers (see Results) were discarded. Its equation is $y = 0.25 + 0.74x$, with a correlation factor of 0.8. The mutants considered as outliers are indicated as crosses and the other mutants are shown as circles.

below 5%. The mean experimental destabilisation effect $\Delta\Delta G_{exp}$ of these mutations is 2.46 kcal mol^{-1} and FOLDEF underestimates this effect by, on average, 0.65 kcal mol^{-1}. These values are far from the corresponding values for V54T, $\Delta\Delta G_{exp}$ of 4.8 kcal mol^{-1} and error of 3 kcal mol^{-1}. This means that the underestimation of the solvation penalty associated with the burial of polar atoms cannot account for the observed discrepancy. For the two last outliers D71A and D54A, close examination of the wild-type (WT) structures does not provide any clues about the reason for the large destabilisation observed for these mutations.

Blind test of FOLDEF on single point mutations

Equation (1) describes the relationships between ΔG and the different energy terms taken into account in FOLDEF. It contains several adjustable parameters and weights that govern the relative contribution of these energy terms in the calculation of the protein stability. In the previous section, we used an initial training database of 339 mutants to obtain a set of optimal weights that best fitted the experimental $\Delta\Delta G$. Since the parameters and weights of FOLDEF were fitted on a particular database, several blind tests were made to check the absence of bias towards the training mutant database. A first blind test was carried out on a new mutant database containing 625 experimental $\Delta\Delta G$ values measured for 27 proteins. As in the previous calculation, only mutants involving deletions or substitutions were considered. These data were recovered from the ProTherm database,[28] from the set of mutations characterised on the human lysozyme[23] and from protein G and protein L mutation studies.[49,50] Results similar to those obtained with the initial training database were produced with this blind test. The slope of the correlation is 0.64 and the correlation coefficient is 0.73 with a standard deviation of 1.02 kcal mol^{-1}. As in the previous case, less than 5% of the data were outliers (more than 2σ difference between calculated and experimental values). Removal of the outliers (34 mutants) improved the prediction: slope = 0.73, $R = 0.80$ and standard deviation = 0.84 kcal mol^{-1} (Figure 2). The results obtained with the blind test database prove that no significant bias was introduced from the fitting of the weights and parameters to the training database. It shows also that the size of the training database was sufficient to contain examples of most of the interactions that play a role in protein thermodynamics. (Details related to the properties of the mutants of the blind test are provided in Table 2 of the Supplementary Material.)

Analysis of the outliers in the blind test database

For the blind test database, we have checked which factors may explain the discrepancies observed for the outliers (Table 2). It is worth noticing that the majority of the outliers are staphyloccocal nuclease mutants (SNase, PDB code 1STN). An extensive analysis of the thermodynamic properties of the mutants of this protein has been carried out in Shortle's group over the past ten years.[51–53] Many mutants showed large variations in the m_{GuHCl} value and this has been associated with changes in the properties of the

Table 2. Analysis of the outliers in the blind test mutant database

PDB	Mut	ASA (%Acc)	DSSP	$\Delta\Delta G_{exp}$ (kcal mol^{-1})	$\Delta\Delta G_{calc}$ (kcal mol^{-1})	Possible origin of the discrepancy
1BPI	F22A	13.1	S	2.00	4.38	Large hydrophobic buried; possible structural relaxation
1BPI	N44A	14.1	C	3.30	0.47	No interpretation for such a large $\Delta\Delta G_{exp}$
1BPI	R1A	55.3	C	0.5	2.75	Residue at the N terminus. Probably very mobile in solution. The strength of the interactions observed in the structure should be reduced
1BVC	H93G	21.4	H	0.01	2.53	In contact with the ligand in the WT structure. Conformation of H93 in WT misleading
1DYJ	V75A	6.8	S	−0.10	2.1	Large hydrophobic buried; possible structural relaxation
1HFZ	Y103A	19.7	3	2.39	4.69	Large hydrophobic buried; possible structural relaxation
1HGU	S71A	28.3	C	0.97	3.92	N-cap of a helix. Origin of the discrepancy unknown
1IOB	K97G	44.5	3	1.20	−2.98	(NH^{3+}) group pointing into the protein interior. Well predicted using another structure: $\Delta\Delta G_{calc} = 1.27$ kcal mol^{-1} using 2I1B
1QHE	F7L	0.3	S	0.04	2.26	Large hydrophobic buried. Possible structural relaxation
1STN	D83A	24.5	C	3.80	1.37	Slight variation of m ($m = 0.89$). Possible additional problem
1STN	F76G	9.3	S	4.70	2.45	No significant variation of m ($m = 1.05$). Unknown reason
1STN	I72A	8.5	S	5.10	2.70	Mutation affects greatly the unfolded state. $m = 1.29$
1STN	L108A	5.5	T	5.80	2.72	Mutation affects greatly the unfolded state. $m = 0.77$
1STN	L137G	29.6	C	4.60	2.14	Mutation affects greatly the unfolded state. $m = 0.74$
1STN	L38G	6.0	S	0.6	3.71	Large hydrophobic buried. Possible structural relaxation
1STN	M98A	9.7	S	4.60	2.32	Mutation affects greatly the unfolded state. $m = 0.75$
1STN	M98G	9.7	S	4.50	2.01	Mutation affects greatly the unfolded state. $m = 0.8$
1STN	N100A	0.0	H	5.20	0.97	Mutation affects greatly the unfolded state. $m = 0.8$
1STN	N100G	0.0	H	5.10	2.96	Mutation affects greatly the unfolded state. $m = 0.71$
1STN	N138G	29.2	C	−0.1	2.1	Mutation affects greatly the unfolded state. $m = 0.87$
1STN	P117A	42.6	T	−0.80	1.36	*cis*-Proline. Turn VIA may convert to turn I (no variation of $m = 1.04$)
1STN	P117G	42.6	T	−0.90	1.81	*cis*-Proline. Turn VIA may convert to turn I (no variation of $m = 0.94$)
1STN	P42G	11.2	C	0.40	2.69	Slight variation of m ($m = 0.89$). Possible structural relaxation
1STN	R105A	32.9	H	1.40	−0.88	No significant variation of m ($m = 0.97$). No interpretation
1STN	V111A	10.4	S	4.20	1.70	Mutation affects the unfolded state. $m = 0.64$
1STN	V111G	10.4	S	4.90	2.31	Mutation affects the unfolded state. $m = 0.75$
1STN	V23G	0.7	S	5.60	3.11	Mutation affects the unfolded state. $m = 1.34$
1STN	V99T	0.7	H	3.30	1.00	No significant variation of m ($m = 1.07$). No interpretation
1WSY	E49Q	1.6	S	2.50	−1.03	Discrepancy not understood. E49A and E49G are well predicted
1WSY	P57A	13.2	C	0.48	3.06	In flexible loop, exposed. P57 may be exposed in solution
2LZM	L99G	0.0	H	6.30	3.94	L99G results in a 4–5 Å displacement of part of helix a solvent-accessible declivity[92]
1REX	I106A	4.3	3	0.93	3.75	Large hydrophobic buried; possible structural relaxation
1REX	S24A	40.6	C	0.53	2.68	Ser at helix N-cap. The C=O_{24} (backbone) and not the side-chain is capping the amide NH27. The N-cap penalty (see Materials and Methods) is applied in a wrong case
1REX	T43V	45.3	S	−0.96	1.19	V43 changes conformation to interact with L85, as seen in the mutant crystal structure[93]

denatured state. Direct evidence of the native like properties of the denatured state of SNase, even in 8 M urea, was reported recently.[54] The variations of m_{GuHCl} values, from 10 to 36% of the WT value, for 13 outliers of SNase, suggest that the effect of the mutations on the properties of the denatured state is responsible for the large discrepancy between $\Delta\Delta G_{exp}$ and $\Delta\Delta G_{calc}$. Although such pronounced mutant effects are specific to the SNase protein, they have been observed in other proteins.[55] Yet, no systematic underestimation of the $\Delta\Delta G_{exp}$ could be observed for the other proteins tested. Overall, among the 34 outliers identified, a likely explanation for the discrepancy can be proposed for 27 mutants (79%) (Table 2). Large discrepancies may arise either from structural relaxation as observed directly in some mutant structures (L99G in 2LZM or T43V in 1REX), from large variations in the m_{GuHCl} values (11 cases), from the existence of contacts with ligand (H93G in 1BVC) and, finally, through crystal packing effects (K97G in 1IOB) (Table 2).

Prediction of stabilising mutations

In the training database and in the blind test databases, most of the mutations are destabilising. To verify that FOLDEF parameters are suitable to estimate the effect of stabilising mutations, we considered a set of 42 mutants of the T4 lysozyme (X → A or G) whose structures have been solved by X-ray studies.[35] The correlation between $\Delta\Delta G_{exp}$ and $\Delta\Delta G_{calc}$ for these mutants is quite high, with a slope of 0.87, a correlation coefficient of $R = 0.83$; standard deviation = 0.89 kcal mol^{-1} (Figure 3). On the basis of the mutant structures, we modelled the reverse mutations (A or G → X) using the WHAT IF program[56] (see Materials and Methods) and calculated $\Delta\Delta G$ for the reverse mutation. We get a very good correlation, with a slope of 0.92, a correlation coefficient of 0.81 and a standard deviation of 0.98 kcal mol^{-1} (Figure 3). All together, the 84 destabilising and stabilising mutations made on the phage T4 lysozyme are predicted with a correlation of 0.89 and a slope of 0.86; standard

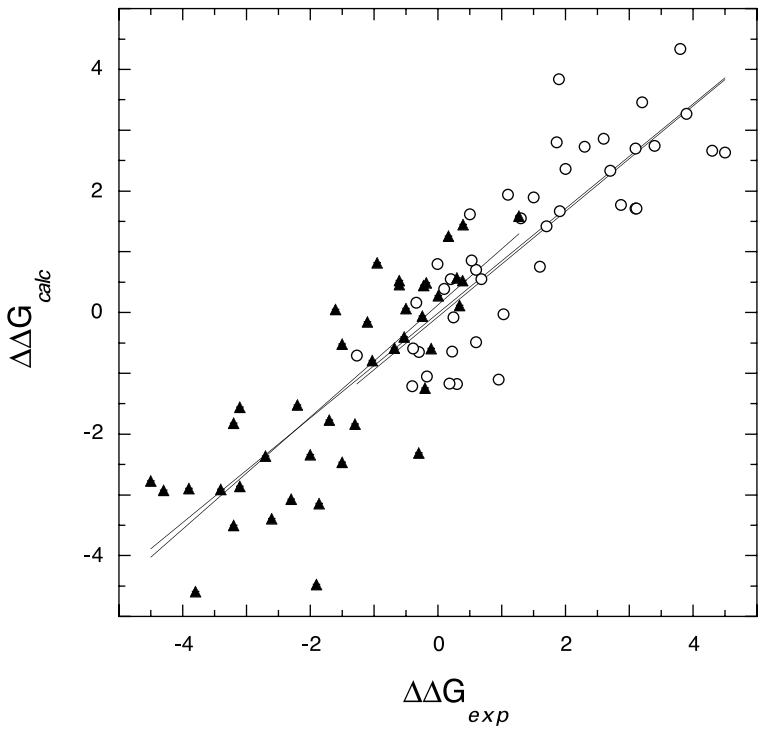

Figure 3. Calculated ΔΔ*G*s compared to the experimental ΔΔ*G*s for the 84 mutants of the T4 lysozyme database. The continuous lines represent the linear regressions obtained considering the 42 mutations (X → A, G) shown as open circle (eq: $y = -0.072 + 0.87x$, $R = 0.83$), the 42 reverse mutations (A, G → X) shown as filled triangles (eq: $y = 0.12 + 0.92x$, $R = 0.81$), or the sum of them (eq: $y = -0.013 + 0.86x$, $R = 0.81$), or the sum of them (eq: $y = -0.013 + 0.86x$, $R = 0.89$).

deviation of 0.97 kcal mol^{-1}. These results demonstrate the ability of FOLDEF to predict the effect of stabilising mutations accurately (Figure 3). It should be emphasised that a critical step of this prediction is the correct modelling of the mutated side-chain. Here, because the crystallographic structures of the mutant and of the WT proteins are very similar, the modelling of the mutated side-chain is relatively straightforward and consequently we find the calculations to be accurate.

Protein–protein complexes

A final set of blind energy predictions were carried out to investigate the possibility of using FOLDEF to estimate the variation in binding free

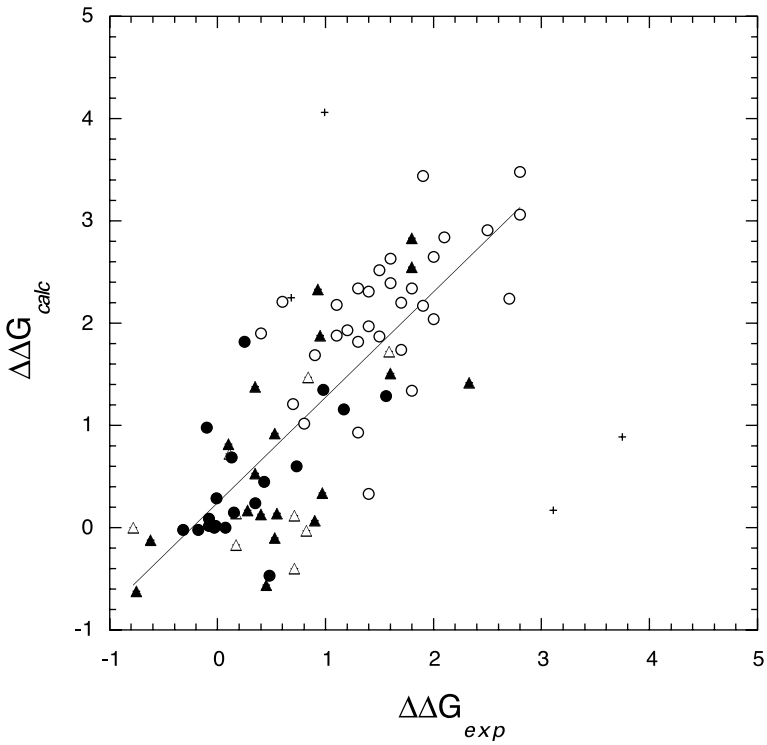

Figure 4. Calculated ΔΔ*G*s compared to the experimental ΔΔ*G*s for the 82 mutants of the protein–protein complex database. The continuous line represents the linear regression obtained with 95% of the training database after the outliers (see Results) were discarded. The corresponding equation is $y = 0.24 + 1x$ with a correlation factor of $R = 0.8$. The symbols used are the crosses for the outliers, the open triangles for the SH3–ligand mutants, the filled circles for the IL4/IL4 receptor mutants, the filled triangles for the P53 tetramer mutants and the open circles for the TEM–BLIP complex mutants.

Table 3. Summary of the results

Database	No. mutants	Correlation	Standard deviation (kcal mol^{-1})	Slope
Initial database	323 (339)	0.8 (0.7)	0.75 (0.97)	0.73 (0.67)
Blind test database	591 (625)	0.8 (0.73)	0.84 (1.02)	0.73 (0.64)
T4 lysozyme mutants (X → A,G)	42	0.83	0.89	0.87
T4 lysozyme mutants (A,G → X)	42	0.81	0.98	0.92
T4 lysozyme mutants (all)	84	0.89	0.97	0.86
Initial + blind test + T4 (A → X)	952 (1006)	0.83 (0.77)	0.81 (1.00)	0.76 (0.69)
Protein–protein complex	78 (82)	0.8 (0.64)	0.66 (0.88)	1.03 (0.79)
Entire mutant database	1030 (1088)	0.83 (0.75)	0.81 (1.00)	0.76 (0.69)

Values in parentheses correspond to the results obtained considering all the mutants of a given database, and values without parentheses are those obtained considering 95% of the mutants of the database.

energy ($\Delta\Delta G_{KD}$) due to mutations made at the interface of protein–protein complexes. We calculated binding energies for four well-studied protein–protein complexes. X-ray structures are available for all of these complexes and extensive mutagenesis experiments at the protein–protein interface for each of them allow a statistical estimation of the quality of the energy function. We investigated the ability of FOLDEF to reproduce complex cooperative hydrogen-bond network phenomena by studying mutants of the TEM–BLIP complex.[57] The ability of FOLDEF to predict the variation of binding free energies $\Delta\Delta G_{KD}$ upon mutation for several SH3 ligands was investigated using the Abl-SH3 domain, and the experimental data obtained from Ala scans of the p53 tetramer and of the IL4–IL4 receptor complex further demonstrated the ability of FOLDEF to predict accurately the change in binding free energies upon mutation of a protein–protein complex.

The global correlation obtained for the entire set of 82 mutants between the experimental and the theoretical $\Delta\Delta G_{KD}$ is shown in Figure 4. Overall, we obtain a correlation of 0.64 with a standard deviation of 0.88 kcal mol^{-1}. When the four outliers showing large discrepancies (larger than 2σ) between $\Delta\Delta G_{exp}$ and $\Delta\Delta G_{calc}$ are excluded, the correlation rises to a larger value, $R = 0.8$ with a standard deviation of 0.66 kcal mol^{-1}, which is the lowest standard deviation obtained in our analysis so far. One of the outliers corresponds to the prediction of SH3-interaction with ligand p28, two others to the IL4 receptor experiments (E9Q and R88A mutations), and the last corresponds to the p53 tetramer experiment (A347G mutation). For the p28 peptide, the side-chain of F59 could not be modelled without introducing a clash with H32 or with W36. This indicates that the structure of the mutant complex should re-arrange upon mutation, and this feature is not considered in the present modelling procedure. We have no clear explanation for the discrepancy observed for the two mutants E9Q and R88A in the IL4 complex experiment, except that the experimental $\Delta\Delta G_{KD}$ values are very high (greater than 3 kcal mol^{-1}), indicating a probable change in the structure. Finally, the A347G mutation creates a cavity at the interface between the helices interacting in the p53 tetramer and the effect of this mutation is overestimated in FOLDEF calculation. A structural rearrangement of the tetramer may occur upon mutation involving tighter packing of the helices when a Gly residue is at the interface. (The detailed values obtained together with the experimental data are presented in Table 2c of the Supplementary Material.)

The results obtained with the protein–protein complexes show that the principles and parameters used in FOLDEF can apply to folding free energy calculations and to binding free energy calculations. The FOLDEF algorithm may thus be used as a tool to guide the engineering of the protein–protein complex interface. Given the rapid time of calculation of FOLDEF, it may prove to be a useful tool for protein–protein docking analysis.

Summary of the results

The predictions of FOLDEF presented here have been carried out on the largest mutant database ever tested (see summary in Table 3). First, we used a set of 339 mutants to adjust the weights of the energy terms in the energy function. For 95% of the mutants, we obtained a correlation of 0.81 between $\Delta\Delta G_{calc}$ and $\Delta\Delta G_{exp}$, and a standard deviation of 0.75 kcal mol^{-1}. We then tested the predictive power of FOLDEF on a blind test database containing 625 mutants. The results obtained with the blind test database are very similar to those obtained with the training database, with a correlation of 0.80 and a standard deviation of 0.84 kcal mol^{-1}. This confirmed that the fitting procedure did not introduce any bias towards the training database. Using a set of mutants of T4 lysozyme, we showed that the potential accounts for the destabilising effect of mutations and for stabilising effects (correlation of 0.81 for the prediction of the stabilising mutations). Finally, we expanded the test of the energy function to include mutations at the interface of protein–protein complexes. We showed that the parameters used in FOLDEF to predict the change of free energy of unfolding upon mutation had the same accuracy for predicting the change of binding free energy (Table 3). The global correlation obtained for 95% of the mutant database (952 mutants) is 0.83 with

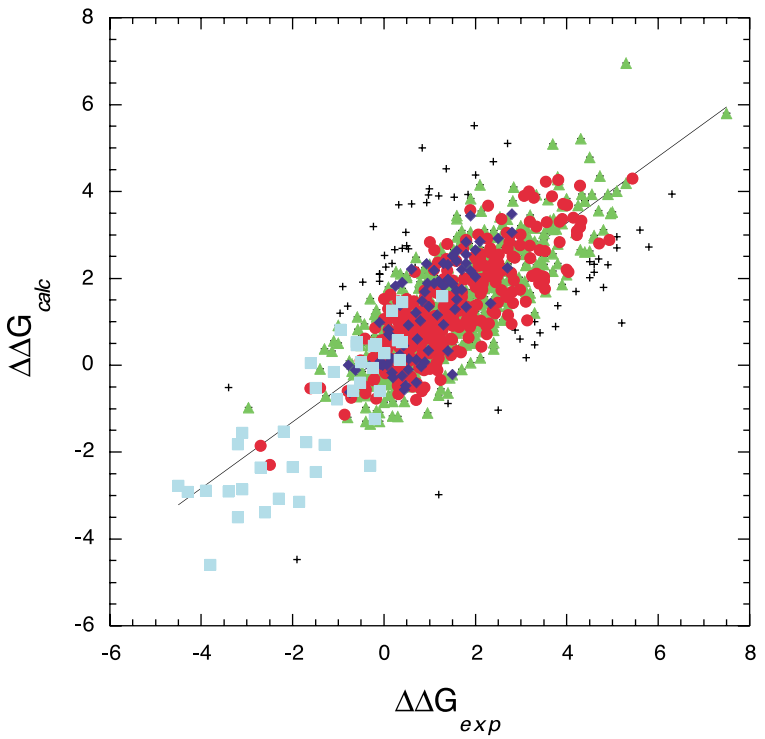

Figure 5. General presentation of the calculated ΔΔ*G*s compared to the experimental ΔΔ*G*s for the 1033 mutants considered in the study. The continuous line represents the linear regression obtained with 95% of the training database after the outliers (see Results) were discarded. The corresponding equation is $y = 0.22 + 0.76x$ with a correlation factor of $R = 0.83$. The symbols used are the crosses for the outliers, the red circles for the training database mutants, the green triangles for the blind test database, the cyan squares for the T4 lysozyme mutants (A, G → X), and the blue diamonds for the protein–protein complex mutant database.

a standard deviation of 0.81 kcal mol^{-1} and a slope of 0.76. We got identical results after adding the mutants of the protein–protein complex database (1030 mutants in total) (Table 3 and Figure 5).

Discussion

The strategy used in this work is based on the large number of protein mutants whose thermodynamic properties have been studied experimentally. Hence, the FOLDEF energy function includes the energy data derived from model compound studies, and accounts for the features specific to the protein world. These features are, for instance, the importance of the structural flexibility, the existence of the unfolded state as a reference state, and the dielectric properties of the protein in the core or at the surface. In all these cases, the strength of an interaction is dependent on the structural context. FOLDEF is designed to integrate the structural environment of an interaction and to predict the impact of this environment on the interactions. The specific features taken into account in FOLDEF are discussed below.

Protein flexibility

The principle of FOLDEF is to take into account implicitly the variation of flexibility in different regions of the protein. It was highlighted recently that the packing density around each atom is a suitable parameter to predict the flexibility in proteins.[58] In fact, a related parameter (number of contacts around an atom) was shown to improve the prediction of point mutations in different proteins significantly.[34,59] FOLDEF calculation is based on the atom occupancy parameter that reflects directly the packing density around atoms. This is one of the reasons why this parameter was preferred to the ΔASA parameter, in addition to the fact that it is much faster to calculate.[36] In other methods that we discuss later, the use of the ΔASA parameter was not sufficient to account for flexibility mutants.[60] The knowledge of the experimental *B*-factors was required and added to the equation. In FOLDEF, the effect of the flexibility is implicitly predicted by the occupancy of each atom. The local flexibility of the protein is taken into account also in the way the side-chain and backbone entropy penalties are applied. Independent of whether side-chains are hydrogen bonded, and whether the backbone is involved in a hydrogen bond network, the entropy penalties are either fully applied or scaled down (see Materials and Methods). All these considerations are crucial to account for protein flexibility and for their thermodynamic properties. The way they are integrated at the different stages of a FOLDEF calculation partly explains the success of the algorithm.

Secondary structure propensities

Several experimental studies have highlighted the role of secondary structure propensities in protein stability. On the one hand, introducing residues that have favourable secondary structure propensities at certain positions in a protein can produce significant increases in protein stability.[24,26,61,62] On the other hand, the opposite result is found, as shown, for instance, in the protein GB1. Several mutants were designed in

the α-helix of this protein so that the propensity for forming β-hairpin structure was increased. Although this variation of intrinsic propensity did not prevent the formation of the native structure, it had an important destabilisation effect on the protein.

To account for this effect, which is related to the properties of the unfolded state, we have included secondary structure propensities in the main-chain entropy term. On the basis of the statistical analysis of the PDB, this term provides a rough estimation of the intrinsic preference of each amino acid to adopt particular φ/ψ angles. As an example, part of the destabilisation due to the mutation of an alanine residue into a glycine residue is accounted for by the main-chain entropy term. In future developments it may be possible to add more information, such as the secondary structure of the neighbouring residues. Hence, the effect of mutations in capping, turns or other constrained local structures could be better predicted. One of the difficulties is to prevent overlaps between the main-chain entropy term and other energetic factors.

The unfolded state as a reference state

The properties of the unfolded state are implicitly taken into account in FOLDEF through the optimal values of the weights obtained in equation (1). These weights have been obtained from the fitting of the training mutant database. The values of the optimal weights are ($W_{vdw} = 0.2$, $W_{solvH} = 1.4$, $W_{solvP} = 1.25$, $W_{sc} = 0.75$ and $\Delta G_{hbond} = -1.3$ kcal mol^{-1}). Since the mutants of the blind test mutant database were equally well predicted, we conclude that no bias was introduced from the first fitting procedure and we can discuss the physical meaning of the weights obtained.

Regarding the weight applied to the side-chain entropy (W_{sc}), we observed from the grid search procedure that values below 1 always gave the best correlation with the experimental data. This could be due to the fact that, in the unfolded state, the side-chains do not adopt all their possible rotamers due, for instance, to the existence of residual structure[54] or neighbouring residues. With respect to the theoretical value of the side-chain entropy,[54] a decrease by 75% ($W_{sc} = 0.75$) was found optimal to account for the experimental measures.

Regarding the weight applied to the van der Waals interactions (W_{vdw}), all the combinations tested in the grid search procedure showed that the best correlations were obtained with W_{vdw} ranging from 0.2 to 0.4. The initial values of the van der Waals energies, before they are weighted, are derived from the transfer of model compounds from vapour to water.[63] In the unfolded state, the polypeptide chain makes substantial van der Waals interactions with the solvent molecules. For that reason, the initial energies have to be decreased. At the extreme, it has been proposed that the van der Waals interactions may not contribute to protein stability, since the interactions with the solvent in the unfolded state would compensate for the interactions made in the folded protein.[64] The results obtained here do not support such a drastic interpretation and show that, although the weight applied to the van der Waals interactions is low, their final contribution to the protein stability after the weight is applied is still important (see below). We found that, on average, the energy of the van der Waals interactions corresponds to two-thirds of that of the hydrophobic solvation effects.

Properties of the protein environment

Regarding the weights applied to the polar and hydrophobic solvation terms, we found that the optimal values were between 1.2 and 1.4. We speculate that these weights reflect the properties of the protein core rather than that of the unfolded state, as in the cases discussed above. The initial values used in FOLDEF (before the weights are applied) were extracted from experiments studying the transfer of model compounds from water to octanol, dioxane or ethanol (see the Supplementary Material for details).[65–68] Solvents such as octanol or ethanol most likely constitute a more polar environment than the core of proteins. This explains why the solvation values of polar and hydrophobic groups have to be increased by 25% or 40%, respectively, to account for the deletion of such chemical groups in the core of protein.

How do the energy terms balance?

The previous discussion highlights the interest in using empirical potentials based on the EEEF approach (as presented in Introduction) to account for the thermodynamic properties of proteins. They allow for an interpretation of the weights used in equation (1) through physical chemistry. The relative strength of the interactions stabilising protein and protein complexes are, on average, well represented by the terms and weights used in FOLDEF. They can therefore be used as a general framework to further understand the thermodynamic properties of proteins. In the following, we present how, on average, the energetic terms compensate each other in the calculation of the energy ΔG of the entire protein. We then discuss their relative importance in the calculation of $\Delta\Delta G$ upon mutation.

The mean contribution of each energy terms has been reported for the 42 proteins tested in this study normalised with respect to their number of residues (Figure 6). It can be seen that, overall, the energy terms taken into account in FOLDEF compensate each other well. We found that 31 out of the 42 proteins have a global ΔG between -15 kcal mol^{-1} and 10 kcal mol^{-1}. This is reasonable if one considers the sum of the energy values

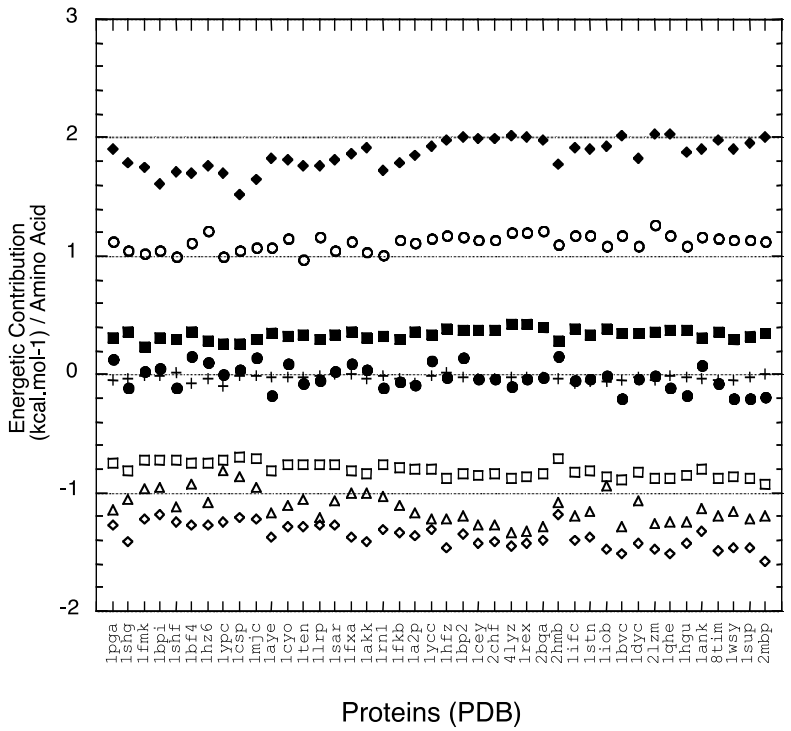

Figure 6. Contributions in kcal mol^{-1} of the different energy terms to the global ΔG calculated for the 42 proteins tested in this study (normalised with respect to the size the protein). The PDB codes of each protein are indicated in the *x*-axis. The different energy terms are the solvation of polar atoms (filled diamond), the solvation of non-polar atoms (open diamond), the hydrogen bonds (open circle), the entropy of the side-chain (filled square), the electrostatic for charge–charge interactions (crosses), and the total free energy (filled circle).

brought by each term (summing in absolute value ~700 kcal mol^{-1} for a 100 amino acid residue protein) and the structural heterogeneity of the tested proteins. It can be seen in Figure 6 that the penalty due to the burial of polar atoms is the major factor opposing the folding of the protein. This effect is not balanced fully by the favourable energies brought by the hydrogen bonding term. The value of the hydrogen bonds, ΔG_{hbond}, has been fit between -1.2 kcal mol^{-1} and -1.4 kcal mol^{-1}, which is in good agreement with the values previously estimated from mutant analysis[34,69,70] and helix/coil approximations.[21,71] Summing the main-chain and side-chain entropy, we found an average value of 1.5 kcal mol^{-1} per residue, which is in good agreement with previously made estimations.[5] It can be noticed that the contribution of the electrostatic interactions, although low compared to the other terms, is in the same range as the final protein ΔG. It indicates that cancelling this contribution may very likely unfold the protein partially, as it can be seen with proteins studied at extreme pH values.

Table 4 shows the average and relative contribution of each energetic factor in the calculation of $\Delta\Delta G$ for a mutation made in different protein environment. For the sake of clarity, we considered here four cases: whether the mutated residue is (i) exposed, (ii) buried (threshold fixed at ΔASA of 30%), (iii) hydrophobic, (iv) polar, involved in at least one hydrogen bond. Compared to the previous paragraph, these results can be seen as the effect of side-chain substitution in protein whereas the previous plot referred to the contribution of the entire backbone and side-chain atoms of the protein. A first comment from this Table is that

Table 4. Average and relative contribution (%) of the different energy terms in the calculation of $\Delta\Delta G$ for different types of mutations

	Type of chemical group mutated			
	Non-polar		Polar involved in at least one H bond	
Energy term	Buried	Exposed	Buried	Exposed
Non-polar solvation	48.4	35.2	15.9	15.7
Polar solvation	11.2	19.7	26.9	23.2
Van der Waals	21.1	18.8	9.2	8.6
Hydrogen bond	–	–	28	25.7
Electrostatic (charge–charge)	–	–	2.8	5.4
Main-chain entropy	7.1	17.9	3.3	4.3
Side-chain entropy	10.5	6	11	15.1

The threshold for deciding that a residue is buried or exposed has been set to 30% for the ΔASA of the mutated residue.

there are no terms negligible in the calculation. In the case of hydrophobic residues, $\Delta\Delta G$ is dominated by hydrophobic and van der Waals energies (70% for buried and 55% for exposed residues). This can explain why a simple method counting the contacts around hydrophobic residues can be quite successful at predicting $\Delta\Delta G$.[34,59] Yet, the results show here that the situation is more complex for polar residues. No individual term dominates, and only a delicate balance between the different terms can yield a proper estimation of $\Delta\Delta G$. For instance, it is interesting to note that for polar mutations, the contribution of the terms favouring the formation of a single hydrogen bond (ΔG_{hbond} and ΔG_{vdw}) almost equal that of the terms opposing its formation (ΔG_{solvP} and ΔS_{sc}). The detailed structural context of the interactions, such as the existence of a hydrogen-bond network and the extent to which the interacting residues are desolvated in the mutant protein, rule the amplitude of the calculated $\Delta\Delta G$ value.

Structural environments such as the existence of water bridges, or the location at the N-cap of helices occurs rarely in our set of mutants and were not considered in the previous analysis. Yet, in specific cases their contribution can be large and should not be ignored (more than 1.5 kcal mol^{-1} for a dozen mutants involving the removal of buried water bridges). We believe that future improvement of the method will lie in the inclusion of such rarely encountered structural contexts (helix dipole effects, π-aromatic interactions) that can be essential for correct prediction of protein thermodynamic properties.

In order to find alternative parameters that may be important for further improvement of the FOLDEF predictions, we tested how energy minimisation and the quality of the template structures may affect the calculations (data not shown). After energy minimisation of the protein structures (using the GROMOS force-field), we observed that the global ΔG calculated with FOLDEF also decreases. However, this energy minimisation did not improve the prediction of $\Delta\Delta G$ for the lysozyme mutants. We also noticed a higher rate of success in the predictions when structures obtained at high resolution (below 1.5 Å) were used. The fact that FOLDEF calculation is sensitive to the quality of the structure constitutes a crucial point in extending its use to the field of structure prediction. Since it includes both enthalpy and entropy energy terms, it may be used as a fast scoring method to rank the large number of structures generated by the structure prediction algorithms.

Comparison with other methods

Here, we briefly discuss our results and compare them to results obtained by other methods developed for similar purposes. We discuss only the methods that deal with a large database of mutants and that consider any type of interactions, hydrophobic and polar. One method is an EEEF method based on the analysis of lysozyme mutants,[60] another is a statistical based method (SEEF).[12,13,72]

Recently, Yutani and co-workers studied the relationships between changes in the stability and the structure of 110 mutants from the human and phage T4 lysozymes. By a least-squares fit of the experimental $\Delta\Delta G$, they derived a unique equation that can represent the thermodynamic properties of proteins. A major difference between FOLDEF and the method described by Yutani is that they use the variation in accessible surface area (ΔASA) to calculate the variation of the solvation for polar and hydrophobic groups, whereas FOLDEF is based entirely on the atomic occupancy parameter. In their first prediction, Yutani and co-workers used the parameters derived from the analysis of 54 mutants of human lysozyme to predict the $\Delta\Delta G$ values for 56 T4 lysozyme mutants. A large standard deviation of 2.39 kcal mol^{-1} was obtained between the estimated and the experimental ΔΔ terms. Excluding atoms having a high *B*-factor in the crystal structure (above 70 Å^2), from the calculation, the standard deviation decreased to 1.82 kcal mol^{-1}. When the parameters of the equation were fit on both the human and the T4 lysozymes (110 mutants with no blind test), excluding atoms with a high *B*-factor, the final standard deviation they got was 1.03 kcal mol^{-1}. This standard deviation is slightly higher than that obtained by FOLDEF considering all the mutants of the database (1088 mutants among which 749 are blind predictions). As we mentioned above, there is no need for the use of the experimental *B*-factors in FOLDEF. This point highlights the advantage of using the atomic occupancy parameter instead of the ΔASA calculation, besides the fact that it is faster to compute. We summarise the other significant differences between the approach taken by Funahashi *et al.*[60] and FOLDEF. First, a specific term accounting for the creation of cavities in the structure was introduced by Funahashi *et al.*[60] In FOLDEF, there is no need for such a term, since the energy penalty brought by the cavity is probably accounted for by the decrease of occupancy of the atoms close to the cavity. Last, the mutations involving electrostatic interactions and steric hindrance are taken into account in the FOLDEF (see Materials and Methods) are, at the present stage, not included in the equation derived by Funahashi *et al.*[60]

A different approach based on the use of statistical potential has been used to predict the variation in stability upon mutation.[12,13,72] It is interesting to compare FOLDEF results and those in this work to highlight the weak and the strong points of each approach. The advantage of statistical potentials lies in two points: (i) for local interactions (important for solvent-exposed residues), the SEEFs implicitly account for a complex ensemble of interactions between one residue and its sequence neighbours. These interactions govern,

for instance, the amino acids' local propensities. (ii) They do not require modelling of the mutant side-chain. In some cases, this may have the advantage to account for the adaptation of the protein structure around the site of the mutation implicitly.

For fully exposed residues with solvent accessibility greater than 50%, Gilis *et al.* obtained very good correlation factors 0.87 and 0.86 for a subset of 90% of their mutant database containing 106 and 150 mutants, respectively. Considering only 90% of our mutant database involving residues having an accessible surface higher than 50% (208 mutants), we obtain a much lower correlation of 0.68, with a standard deviation of 0.51 kcal mol^{-1} and slope of 0.65.

Regarding the residues with surface accessibility below 50% the results of FOLDEF on 95% of the database (718 mutants) are much better than those obtained with the statistical method. The correlation between $\Delta\Delta G_{exp}$ and $\Delta\Delta G_{calc}$ is 0.83, with a standard deviation of 0.85 kcal mol^{-1} and a slope of 0.76. For these types of residues, the method developed by Gilis *et al.* had to consider different weighting factors for different degree of solvent-accessibility. The results of their fitting procedure show good correlation for residues below 20% ($R = 0.80$ for 121 mutants) and between 20% and 40% ($R = 0.82$ for 65 mutants) but failed to predict mutants with accessibility comprised between 40% and 50% (48 mutants, 20% of the database considered).

Therefore, the comparison between the statistical potential[12,13,72] and the methods developed in FOLDEF show an advantage of the statistical methods to account for thermodynamic properties at the surface of proteins based on the local interactions of the residues and on their intrinsic propensity to adopt specific secondary structure. For more buried interactions, the atomic resolution and description of the interactions appears as more potent. These results indicate that an optimal energy function to describe the relative strength of the interactions stabilising protein should benefit from both the statistical information of the protein structure database and from the detailed atomic data. Additional statistical information, particularly about the neighbouring residues in the sequence, will be included in FOLDEF to increase the accuracy of the predictions for fully exposed residues.

Conclusion

FOLDEF was developed to provide a fast and quantitative estimation of the importance of the interactions contributing to the stability of proteins and protein complexes. The predictive power of FOLDEF was tested on a very large set of point mutants spanning most of the structural environments found in proteins. The standard deviations (Table 3) indicate that for 70% of the mutants the error was below 0.81 kcal mol^{-1}. This value provides a confidence interval that can be used to assess the reliability of FOLDEF predictions for protein engineering applications. Based on this large-scale analysis, we found that the successful prediction of protein thermodynamic properties from 3D structures requires two major features: (i) energy terms that take into account the fine details of the structure (hydrogen bonds, water bridges) explicitly; (ii) energy terms that account in an implicit manner for specific properties of proteins such as the flexibility or the existence of the unfolded state.

FOLDEF was validated on monomeric proteins and on protein complexes. Hence, the method may be used to drive the engineering of protein complex interfaces. It may be used to understand and predict the specificity of protein–protein recognition. The use of fast methods to estimate the stability of protein conformations is of significant interest for the improvement of structure prediction methods, in particular in the field of *ab initio* predictions. A critical step of the method is the ability to discriminate between wrong and correct structures in a large ensemble of generated folds. As suggested recently,[73,74] the consideration of more detailed structural interactions based on thermodynamic parameters may provide an important filter to identify the correct folds. In this way, the energies calculated using FOLDEF could serve as a useful indicator of the quality of a structural model.

Finally, the present energy function has been integrated in the program FOLD-X, which is designed to predict folding pathways from 3D structure.[4] We hope to improve the prediction of folding pathways by using the present energy function and thus consider topological constraints and detailed interactions in the calculation of the folding pathway. Work along these lines is ongoing.

The FOLDEF is available *via* a web-interface† and the compiled code can be obtained by sending an email to the authors. For groups interested in further developing the algorithm, the source code can be obtained. The modelled structures and experimental $\Delta\Delta G$ values for the training database, the blind test database and the protein complexes are available at the same web address. In this way, the database used in this study can easily be used to benchmark other algorithms designed to predict stability variations upon mutation.

Materials and Methods

Composition of the potential

FOLD-X energy function (FOLDEF) includes several terms: van der Waals interactions, solvation effects, hydrogen bonds, water bridges, electrostatic and entropy

† http://fold-x.embl-heidelberg.de

effects for the backbone and the side-chain see equation (1) in Results.

Solvent exposure

In FOLDEF, interaction energies are scaled with the solvent accessibility of the atoms involved in the interaction. The solvent accessibility is estimated using the atomic occupancy method (Occ),[36–38] which sums the volumes of the atoms j surrounding a given atom i. The occupancy of an atom i, Occ(i), is calculated using equation (2):

$$\mathrm{Occ}(i) = \sum_{j, d_{ij} \leq 6\ \text{Å}} V_j \, \mathrm{e}^{-\left(\frac{d_{ij}^2}{2\sigma^2}\right)} \qquad (2)$$

where V_j is the fragmental volume of atom j[75] (see Table 3 in the Supplementary Material), d_{ij} is the distance between atoms i and j and $\exp(-d_{ij}^2/2\sigma^2)$ is the envelope function. We used a distance cut off of 6 Å and a distance $\sigma = 3.5$ Å was used because it corresponds roughly to the minimum of the Van der Waals potential for two heavy atoms.[38] Two limit values, $\mathrm{Occ}_{\min}$ and $\mathrm{Occ}_{\max}$, are assigned to each atom type as the reference value for a fully exposed and fully buried atom[37] (see Table 3 in Supplementary Material). The FOLDEF energy evaluation uses these two limit values to scale the strength of the interactions by applying a scaling factor, S_{fact}, calculated using the linear equation (3):

$$S_{\mathrm{fact}}(i) = \frac{\mathrm{Occ}(i) - \mathrm{Occ}_{\min}(t_i)}{\mathrm{Occ}_{\max}(t_i) - \mathrm{Occ}_{\min}(t_i)} \qquad (3)$$

if $\mathrm{Occ}(i) > \mathrm{Occ}_{\max}(t_i)$, $S_{\mathrm{fact}}(i) = 1$ and if $\mathrm{Occ}(i) < \mathrm{Occ}_{\min}(t_i)$, $S_{\mathrm{fact}}(i) = 0$.

For every atom type $\mathrm{Occ}_{\min}$ and $\mathrm{Occ}_{\max}$ have been derived from a statistical analysis of a protein structure database.[37] (see Table 3 in Supplementary Material).

van der Waals and solvation energies

Both the van der Waals and the solvation energies were obtained from the free energy of transfer of the amino acids from vapour to water and from organic solvents to water[76] (Table 4 in Supplementary Material). The van der Waals atomic energies (see Table 3 in Supplementary Material) were taken from the free energy of transfer of small compounds from vapour to water (giving the sum of the solvation and the van der Waals energies) and from water to cyclohexane (giving the solvation energies).[63] As proposed,[63] the van der Waals energy have been deduced from the difference between these two sets of parameters. We decomposed the van der Waals energies of every amino acid into atomic energies assuming that the van der Waals energies vary linearly with the fragmental volumes of the atoms as defined in.[75] Using the CH_3 of the Alanine as a reference, we estimated that the van der Waals energy was -0.082 kcal mol^{-1} per unity of atomic volume buried (in Å^3).

The atomic solvation energies were obtained using the same method as for the van der Waals energies. The only difference is that the experimental data were obtained by averaging the values from three different studies. Two of these studies measured the free energies of transfer of amino acids from water to n-octanol[65] and from water to ethanol or dioxane.[66,67] The third one comes from the partition coefficients between water and n-octanol of many model compounds.[68] The three scales correlate well with each other ($R = 0.94$ on average). From these values, we estimated the solvation energy at -0.0314 kcal mol^{-1} Å^{-3} for hydrophobic atoms, at 0.8 kcal mol^{-1} for polar atoms and 1.44 kcal mol^{-1} for charged atoms (Table 3 in Supplementary Material).

We checked that, for the 20 amino acids, the sum of the atomic energies obtained by these methods correlate well with the experimental reference data (correlation factor $R = 0.91$ for the van der Waals energies and $R = 0.99$ for the solvation energies (see Figure 1 in Supplementary Material)).

In the case of polar atoms when the maximum number of hydrogen bonds for a particular atom is reached (or achieved), we assumed that the atom is completely desolvated and apply the maximum solvation penalty regardless of the occupancy value (Occ(i)).

Hydrogen bonds

This calculation concerns polar/polar and polar/charged atoms pairs are within 3.6 Å. A hydrogen bond between two atoms is accepted or rejected based on specific angle criteria.[77] The contribution of a hydrogen bond to the free energy has been estimated at -1.3 kcal mol^{-1} between two polar atoms and -1.4 kcal mol^{-1} between a charged and a polar atom. These particular values result from the fitting of the training database (Figure 1) and are consistent with the range of values usually proposed for hydrogen bonds.[34,69]

Electrostatics

Electrostatic energies are calculated between charged atoms of the N and C termini, and between the charged atoms of Asp, Glu, Arg, Lys and His residues only if they are closer than 20 Å. The contribution of the electrostatic interaction to the free energy calculation is determined using Coulomb's equation with an ionic strength screening term (4):

$$E_{ij} = \frac{332 q_i q_j}{\varepsilon d_{ij}} \exp(-d_{ij} K) \ \ (\text{in kcal mol}^{-1}) \qquad (4)$$

where q_i and q_j are the charges of atoms i and j, as defined in Table 3 in Supplementary Material, ε is the dielectric constant of the medium, d_{ij} is the inter-atomic distance between i and j, and K is the Debye–Hückel parameter to account for ionic strength effect of the solution defined as:

$$K = (8\pi e^2 N I / 1000 kT)^{1/2} = 5.66 \sqrt{\frac{I}{T}} \ \ (\text{in Å}^{-1}) \qquad (5)$$

where I is the ionic strength of the solution (in M), N is the Avogadro's number, k is the Boltzmann's constant and T is the temperature (in K).

All the calculations were done with ε linearly increasing from 8 to 80 with the scaling factor S_{fact} (see above), $T = 298$ K and $I = 0.05$ M. As a first approximation, we assumed that all amino acid pKa values were unperturbed by the protein environment, and correspondingly we assigned the standard charge at pH 7.0 to all residues.

Backbone and side-chain entropy

The backbone entropy term is used to account for the entropy cost of fixing a residue backbone.[4] The value of

the backbone entropy was calculated from the secondary structure preference of amino acid derived from the statistical analysis of a protein structure database.[32] These values were scaled so that values for Ala, Gly and Val residues in helix and strand conformations are the same as those used.[4] For that reason, the values of the main chain entropy were not further adjusted and the weight of the main chain entropy W_{mc} was maintained at 1.

The backbone of residues located in loops is usually more mobile than in secondary structure elements and the backbone entropy should therefore not be counted fully in these cases. To account for this effect we applied a simple rule for residues forming any backbone–backbone hydrogen bond: if none of the two neighbouring residues are involved in backbone–backbone hydrogen bonds, the backbone entropy is divided by 3. If only one of the neighbouring residues forms no backbone–backbone hydrogen bonds, the backbone entropy is divided by 2. Regarding Pro, it is well known that a non-Gly residue preceding a Pro residue can mainly adopt an extended conformation and thus has less conformational freedom in the unfolded state.[78] Thus, the main chain entropy of non-Gly residues preceding Pro is divided by 2.

The side-chain entropy is calculated from the values estimated.[33] The scaling factor S_{fact} is applied to the side-chain entropy term to account for the fact that the mobility of a side-chain decreases with its solvent accessibility. However, if a side-chain makes a hydrogen bond or an electrostatic interaction, then we apply the full entropy cost since the formation of an interaction reduces the mobility of the side-chain. If the entropy cost is bigger than the favourable interaction energy brought by the hydrogen bond and the electrostatic interactions, neither these interactions nor the entropy of the side-chain are counted.

Water bridges

The calculation of the effect of water in protein stability is a complex issue. Several experimental studies show that the deletion of polar atoms that make hydrogen bonds with a partially of fully buried water molecule can have a large destabilising effect on the protein interaction.[30,31,79,80] We define a water bridge as a water molecule that makes more than two hydrogen bonds with the protein. Removing one of the polar groups involved in a water bridge may exclude the bound water from a particular site of the protein and induce the desolvation of the other polar groups partners of the water molecule.

In the FOLDEF, the energy assigned to a water bridge interaction allows us to reduce the solvation penalty for buried polar atoms when they are involved in such an interaction. To predict the positions of water bridges, we used the method described.[81] This method considers all the potential locations of water molecules for all polar atoms. Water molecules that make van der Waals clashes with the protein atoms are discarded (limit distances of 2.6 Å and 3.1 Å for N and for O, C and S atoms, respectively[82]). We predict the existence of a water bridge when two water molecules are found within a 2.8 Å of each other. The two molecules are then fused and the coordinates of a mean water molecule are calculated. When a likely water bridge is found, a search for other polar groups that are likely to contribute to the hydrogen bond network around the water molecule is performed. The following fast calculation is used to estimate ΔG_{wb}, the contribution of a water bridge to the stability of a protein using equation (6). Only the water bridges with negative ΔG_{wb} are added to the global energy of the protein.

$$\Delta G_{wb} = N_{hb}\Delta G_{hb} + S_{fact}\Delta G_{solvW} + \delta S_{prot} + (1 - S_{fact})S^{max}_{wat} + S_{fact}S^{min}_{wat} \quad (6)$$

where N_{hb} is the number of hydrogen bonds between the water and the protein, ΔG_{hb} is the energy of a hydrogen bond, ΔG_{solvW} is the solvation cost for water burial (see Table 3 in Supplementary Material), S_{fact} is the scaling factor of the water molecule and δS_{prot} is the additional entropy cost associated with fixing the side-chain or the main-chain involved in the water bridge. S^{min}_{wat} and S^{max}_{wat} are the entropy penalties associated with the fixation of a water in a fully buried and fully solvent exposed position, respectively.

At the surface of a protein, a water molecule can adopt many more configurations than when it is fixed in a cavity.[31] Hence, the fixation of the water molecule in the configuration where it makes the full water bridge should have a higher entropy penalty at the surface than in the core of the protein. It is important to note that the term S^{max}_{wat} may also reflect additional factors that reduce stability of water bridges in fully exposed location. We used the entropy cost for fixing a buried water molecule $S^{min}_{wat} = 0.92$ kcal mol^{-1}[83] and $S^{max}_{wat} = 2.5$ kcal mol^{-1} at room temperature.

Additional features taken into account in the potential

van der Waals clashes

In some structures of the protein database, van der Waals clashes are observed and can be due to the resolution of the structures. Given that FOLDEF is based on the atomic occupancy, the van der Waals clashes result in an overestimation of the solvation and van der Waals energies. To circumvent this problem a term has been introduced to account for van der Waals clashes in proteins. The condition for the existence of a clash is that $d_{ij} < (R_i + R_j - 0.35)$. The value of 0.35 Å is a tolerance factor, which corresponds to the resolution of the crystal structure (typically <2 Å). d_{ij} is the interatomic distance between atoms i and j and R_i and R_j are the corresponding atom radii given in the Table 1 of Supplementary Material. The correction in energy is given by the formula:

$$\Delta G_{clash} = \Delta_{clash} S_{fact} \quad (7)$$

where

$$\Delta_{clash} = R_i + R_j - 0.35 - d_{ij} \quad (8)$$

and S_{fact} is the scaling factor that takes solvent exposure into account (see above). (For structures with resolution lower than 2 Å, ΔG_{clash} is usually zero and does not excess 1.0 kcal mol^{-1} for one residue in a protein).

N-cap of α-helices

At the N terminus of a α-helix, three amide protons (NH) of the backbone are structurally constrained, extending outwards and close to each other in space. Because of steric constraints, it has been shown that

change in the location of the water molecules bound to these NH could highly affect the stability of the protein.[84] A Gly at the N-cap involves no steric constraints and water molecules can solvate perfectly the amide protons, whereas the Ala side-chain at the same position affects the solvation. The S_{fact} scaling factor, used to estimate the desolvation of the NH groups, is not sensitive enough (no directionality) to take this effect into account. Therefore, if the N-cap residue is neither a Gly nor has a short polar side-chain able to cap the helix N terminus (Asp, Thr, Ser, Asn), a desolvation penalty of 1.5 kcal mol^{-1} is given to the residue at the N-cap. This value was obtained from the analysis of several Ala-Gly mutations made at the N-cap of helices.[34]

Selection of the mutants in the database

The training database contains 339 mutants that were experimentally studied in nine different proteins, Barnase, CI-2, Spectrin and Src SH3, Sso7d, Tenascin, FKBP, Ada2h and CheY, names with ref (see Table 1 in Supplementary Material). The blind test database was built considering all the mutants of the ProTherm database[28]† involving a single conservative mutation made in a monomeric protein and studied between pH 6 and 8. Because they represent a large set of data, we also included the ensemble of mutants of the T4 lysozyme studied at pH 3 or 2. We also included the set of mutations characterised on the human lysozyme[23] and on the protein G and protein L[49,50] that were not yet taken into account in the ProTherm database[28] (see details in Table 2a in Supplementary Material). Data and PDB codes for the T4 lysozyme mutants were retrieved from the ProTherm database[28] (see details in Table 2b in Supplementary Material). The thermodynamic data used in the protein–protein complex database were recovered from Ref. 57 for TEM–BLIP (PDB: 1JTD), Ref. 85 for the SH3–ligand (PDB: 3BP2), Ref. 86 for the P53 tetramer (PDB: 1AIE) and Refs. 87,88 for the IL-4/IL-4 receptor (PDB: 1IAR) (see details from Table 2c in Supplementary Material). In the case of the SH3–ligand mutants, we only considered the ligands whose affinity have been determined using the precise titration method and not using the extrapolation method.[85] For all the complexes, except for the P53 tetramer mutants, the $\Delta\Delta G_{KD}$ were computed by calculating the difference between the energy of the bound state and the energy of the isolated monomers. For P53, since the monomer is unfolded, we only considered the energy of the tetramer. The values of the $\Delta\Delta G$ calculated between the WT and the mutants were divided by 4 to account for the effect of the mutation in one monomer.

Modelling mutants

Point mutations were modelled using a modified version of the WHAT IF[56] mutate functions. The modifications ensured that the C^{β} atoms of the mutated residue(s) were at identical positions in the wild-type and mutant enzymes. We did not change any atom positions for mutations that involved only deletion of atoms (all Ala and Gly mutations, Ile to Val, Tyr → Phe, etc.) or for mutations that only involved changing atom types (Cys → Ser, Val → Thr, Thr → Val, etc.). For the mutations of the T4 lysozyme (A,G → X) and of the SH3–ligand complex, that do not fall in one of the above categories, we used the so-called "experimental" version of the WHAT IF mutate function[89] followed by a debumping step (the WHAT IF DEBHBO function) that optimises the number of hydrogen-bonds for the mutated residue. We did not at any point change coordinates for any other atoms than those in the mutated residue. For the T4 lysozyme (A,G → X) mutants, the van der Waals clash energies were not taken into account in the final energy because steric hindrance of the newly introduced residue with the protein could not be avoided due to the simplicity of the modelling procedure.

Following mutation we optimised the hydrogen-bond network in the protein using a method developed by Hooft *et al.*,[90] since it has been shown[91] that inconsistencies in the hydrogen-bond network (especially His, Asn and Gln residues with wrong X1, X1 and X2 angles) can introduce significant errors in protein energy calculations. The final structures used with FOLDEF consist of all heavy atoms of the protein and the backbone amide protons. All ions and water molecules were stripped from the PDB files before mutating any residue.

Fitting of the weights

The optimal set of weights used in equation (1) were obtained by a grid search method considering first the weights of the van der Waals (W_{vdw}), the solvation of hydrophobic atoms (W_{solvH}) and the side-chain entropy (W_{sc}). We first considered the ensemble of 151 mutations from the training database that involves only hydrophobic residues not identified as outliers. In this way, we excluded electrostatics and hydrogen bond energies from the initial fitting procedure. All combinations of values for the weights W_{vdw}, W_{solvH}, and W_{sc} were tested between 0 and 2 by steps of 0.2. The ten best combinations of weights, W_{bests}, giving the lowest standard deviation error between $\Delta\Delta G_{exp}$ and $\Delta\Delta G_{calc}$, were selected for a second round of fitting on the entire training database. In this second round we considered the weights for the solvation of polar atoms W_{solvP} and the hydrogen bond value ΔG_{hbond}. Every of the W_{bests} combination was tested with W_{solvP} and ΔG_{hbond} varying between 0 and 2 by step of 0.2 and from −1 to −2 by steps of 0.1, respectively. The lowest standard deviation associated with the higher correlation was obtained for the set of values ($W_{vdw} = 0.2$, $W_{solvH} = 1.4$, $W_{solvP} = 1.2$, $W_{sc} = 0.8$ and $\Delta G_{hbond} = -1.4$ kcal mol^{-1}). We further explored improvement of the fit by varying the value of each weight by units of 0.05. The final optimal set of values was $W_{vdw} = 0.2$, $W_{solvH} = 1.4$, $W_{solvP} = 1.25$, $W_{sc} = 0.75$ and $\Delta G_{hbond} = -1.3$ kcal mol^{-1}.

Acknowledgments

We thank N. J. C. Strynadka for giving us the coordinates of the TEM–BLIP structure before release in the PDB (1JTD), and E. Lacroix for providing us with the statistical analysis of the φ/ψ dihedral angles of each amino acid. This work was supported by EU grants BIO4-CT97-2086 and CT96-0013, and by the Ramon Areces Fundation (Spain). R.G. was supported by a fellowship from the Human Frontier Science Program. This work was supported, in part, by NIH and HHMI grants made to J. Andrew McCammon, UCSD. J.E.N.

† http://www.rtc.riken.go.jp

acknowledges support from the Danish Natural Science Council.

References

1. Vajda, S., Sippl, M. & Novotny, J. (1997). Empirical potentials and functions for protein folding and binding. *Curr. Opin. Struct. Biol.* **7**, 222–228.
2. Camacho, C. J. & Vajda, S. (2001). Protein docking along smooth association pathways. *Proc. Natl Acad. Sci. USA*, **21**, 21.
3. Lazaridis, T. & Karplus, M. (2000). Effective energy functions for protein structure prediction. *Curr. Opin. Struct. Biol.* **10**, 139–145.
4. Guerois, R. & Serrano, L. (2000). The SH3-fold family: experimental evidence and prediction of variations in the folding pathways. *J. Mol. Biol.* **304**, 967–982.
5. Alm, E. & Baker, D. (1999). Prediction of protein-folding mechanisms from free-energy landscapes derived from native structures. *Proc. Natl Acad. Sci. USA*, **96**, 11305–11310.
6. Galzitskaya, O. V. & Finkelstein, A. V. (1999). A theoretical search for folding/unfolding nuclei in three-dimensional protein structures. *Proc. Natl Acad. Sci. USA*, **96**, 11299–11304.
7. Munoz, V. & Eaton, W. A. (1999). A simple model for calculating the kinetics of protein folding from three-dimensional structures. *Proc. Natl Acad. Sci. USA*, **96**, 11311–11316.
8. Murphy, K. P. & Freire, E. (1992). Thermodynamics of structural stability and cooperative folding behavior in proteins. *Advan. Protein Chem.* **43**, 313–361.
9. Pace, C. N., Shirley, B. A., McNutt, M. & Gajiwala, K. (1996). Forces contributing to the conformational stability of proteins. *FASEB J.* **10**, 75–83.
10. Sippl, M. J. (1995). Knowledge-based potentials for proteins. *Curr. Opin. Struct. Biol.* **5**, 229–235.
11. Rooman, M. J. & Wodak, S. J. (1995). Are database-derived potentials valid for scoring both forward and inverted protein folding? *Protein Eng.* **8**, 849–858.
12. Gilis, D. & Rooman, M. (1996). Stability changes upon mutation of solvent-accessible residues in proteins evaluated by database-derived potentials. *J. Mol. Biol.* **257**, 1112–1126.
13. Gilis, D. & Rooman, M. (1997). Predicting protein stability changes upon mutation using database-derived potentials: solvent accessibility determines the importance of local versus non-local interactions along the sequence. *J. Mol. Biol.* **272**, 276–290.
14. Topham, C. M., Srinivasan, N. & Blundell, T. L. (1997). Prediction of the stability of protein mutants based on structural environment-dependent amino acid substitution and propensity tables. *Protein Eng.* **10**, 7–21.
15. Bordo, D. & Argos, P. (1991). Suggestions for safe residue substitutions in site-directed mutagenesis. *J. Mol. Biol.* **217**, 721–729.
16. Maxwell, K. L. & Davidson, A. R. (1998). Mutagenesis of a buried polar interaction in an SH3 domain: sequence conservation provides the best prediction of stability effects. *Biochemistry*, **37**, 16172–16182.
17. Larson, S. M. & Davidson, A. R. (2000). The identification of conserved interactions within the SH3 domain by alignment of sequences and structures. *Protein Sci.* **9**, 2170–2180.
18. Prevost, M., Wodak, S. J., Tidor, B. & Karplus, M. (1991). Contribution of the hydrophobic effect to protein stability: analysis based on simulations of the Ile-96-Ala mutation in barnase. *Proc. Natl Acad. Sci. USA*, **88**, 10880–10884.
19. Pitera, J. W. & Kollman, P. A. (2000). Exhaustive mutagenesis in silico: multicoordinate free energy calculations on proteins and peptides. *Proteins: Struct. Funct. Genet.* **41**, 385–397.
20. Kollman, P. A., Massova, I., Reyes, C., Khun, B., Huo, S., Chong, L. *et al.* (2000). Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Accts Chem. Res.* **33**, 889–897.
21. Lacroix, E., Viguera, A. R. & Serrano, L. (1998). Elucidating the folding problem of alpha-helices: local motifs, long-range electrostatics, ionic-strength dependence and prediction of NMR parameters. *J. Mol. Biol.* **284**, 173–191.
22. Munoz, V. & Serrano, L. (1997). Development of the multiple sequence approximation within the AGADIR model of alpha-helix formation: comparison with Zimm–Bragg and Lifson–Roig formalisms. *Biopolymers*, **41**, 495–509.
23. 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.
24. Villegas, V., Viguera, A. R., Aviles, F. X. & Serrano, L. (1996). Stabilization of proteins by rational design of alpha-helix stability using helix/coil transition theory. *Fold. Des.* **1**, 29–34.
25. 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.
26. Taddei, N., Chiti, F., Fiaschi, T., Bucciantini, M., Capanni, C., Stefani, M. *et al.* (2000). Stabilisation of alpha-helices by site-directed mutagenesis reveals the importance of secondary structure in the transition state for acylphosphatase folding. *J. Mol. Biol.* **300**, 633–647.
27. 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.
28. Gromiha, M. M., An, J., Kono, H., Oobatake, M., Uedaira, H., Prabakaran, P. & Sarai, A. (2000). ProTherm, version 2.0: thermodynamic database for proteins and mutants. *Nucl. Acids Res.* **28**, 283–285.
29. Williams, M. A., Goodfellow, J. M. & Thornton, J. M. (1994). Buried waters and internal cavities in monomeric proteins. *Protein Sci.* **3**, 1224–1235.
30. Takano, K., Funahashi, J., Yamagata, Y., Fujii, S. & Yutani, K. (1997). Contribution of water molecules in the interior of a protein to the conformational stability. *J. Mol. Biol.* **274**, 132–142.
31. Xu, J., Baase, W. A., Quillin, M. L., Baldwin, E. P. & Matthews, B. W. (2001). Structural and thermodynamic analysis of the binding of solvent at internal sites in T4 lysozyme. *Protein Sci.* **10**, 1067–1078.
32. Munoz, V. & Serrano, L. (1994). Intrinsic secondary structure propensities of the amino acids, using statistical phi–psi matrices: comparison with experimental scales. *Proteins: Struct. Funct. Genet.* **20**, 301–311.

33. Abagyan, R. & Totrov, M. (1994). Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. *J. Mol. Biol.* **235**, 983–1002.
34. 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.
35. Matthews, B. W. (1995). Studies on protein stability with T4 lysozyme. *Advan. Protein Chem.* **46**, 249–278.
36. Colonna-Cesari, F. & Sander, C. (1990). Excluded volume approximation to protein–solvent interaction. The solvent contact model. *Biophys. J.* **57**, 1103–1107.
37. Holm, L. & Sander, C. (1992). Evaluation of protein models by atomic solvation preference. *J. Mol. Biol.* **225**, 93–105.
38. Stouten, P. F. W., Froemmel, C., Nakamura, H. & Sander, C. (1993). An effective solvation term based on atomic occupancies for use in protein simulations. *Mol. Simul.* **10**, 97–120.
39. Lee, B. & Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **55**, 379–400.
40. Itzhaki, L. S., Otzen, D. E. & Fersht, A. R. (1995). The structure of the transition state for folding of chymotrypsin inhibitor 2 analysed by protein engineering methods: evidence for a nucleation-condensation mechanism for protein folding. *J. Mol. Biol.* **254**, 260–288.
41. Martinez, J. C., Pisabarro, M. T. & Serrano, L. (1998). Obligatory steps in protein folding and the conformational diversity of the transition state. *Nature Struct. Biol.* **5**, 721–729.
42. Riddle, D. S., Grantcharova, V. P., Santiago, J. V., Alm, E., Ruczinski, I. I. & Baker, D. (1999). Experiment and theory highlight role of native state topology in SH3 folding. *Nature Struct. Biol.* **6**, 1016–1024.
43. Hamill, S. J., Steward, A. & Clarke, J. (2000). The folding of an immunoglobulin-like Greek key protein is defined by a common-core nucleus and regions constrained by topology. *J. Mol. Biol.* **297**, 165–178.
44. Fulton, K. F., Main, E. R., Daggett, V. & Jackson, S. E. (1999). Mapping the interactions present in the transition state for unfolding/folding of FKBP12. *J. Mol. Biol.* **291**, 445–461.
45. Villegas, V., Martinez, J. C., Aviles, F. X. & Serrano, L. (1998). Structure of the transition state in the folding process of human procarboxypeptidase A2 activation domain. *J. Mol. Biol.* **283**, 1027–1036.
46. Lopez-Hernandez, E. & Serrano, L. (1996). Structure of the transition state for folding of the 129 aa protein CheY resembles that of a smaller protein CI-2. *Fold. Des.* **1**, 43–55.
47. Consonni, R., Santomo, L., Fusi, P., Tortora, P. & Zetta, L. (1999). A single-point mutation in the extreme heat- and pressure-resistant sso7d protein from sulfolobus solfataricus leads to a major rearrangement of the hydrophobic core. *Biochemistry*, **38**, 12709–12717.
48. Baldwin, E. P., Hajiseyedjavadi, O., Baase, W. A. & Matthews, B. W. (1993). The role of backbone flexibility in the accommodation of variants that repack the core of T4 lysozyme. *Science*, **262**, 1715–1718.
49. McCallister, E. L., Alm, E. & Baker, D. (2000). Critical role of beta-hairpin formation in protein G folding. *Nature Struct. Biol.* **7**, 669–673.
50. Kim, D. E., Fisher, C. & Baker, D. (2000). A breakdown of symmetry in the folding transition state of protein L. *J. Mol. Biol.* **298**, 971–984.
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. 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.
53. 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.
54. Shortle, D. & Ackerman, M. S. (2001). Persistence of native-like topology in a denatured protein in 8 M urea. *Science*, **293**, 487–489.
55. Shortle, D. (1995). Staphylococcal nuclease: a showcase of m-value effects. *Advan. Protein Chem.* **46**, 217–247.
56. Vriend, G. (1990). WHAT IF: a molecular modeling and drug design program. *J. Mol. Graph.* **8**, 29.
57. Albeck, S., Unger, R. & Schreiber, G. (2000). Evaluation of direct and cooperative contributions towards the strength of buried hydrogen bonds and salt bridges. *J. Mol. Biol.* **298**, 503–520.
58. Halle, B. (2002). Flexibility and packing in proteins. *Proc. Natl Acad. Sci. USA*, **99**, 1274–1279.
59. Lopez-Hernandez, E. & Serrano, L. (1995). Empirical correlation for the replacement of Ala by Gly: importance of amino acid secondary intrinsic propensities. *Proteins: Struct. Funct. Genet.* **22**, 340–349.
60. Funahashi, J., Takano, K. & Yutani, K. (2001). Are the parameters of various stabilization factors estimated from mutant human lysozymes compatible with other proteins? *Protein Eng.* **14**, 127–134.
61. Munoz, V., Lopez, E. M., Jager, M. & Serrano, L. (1994). Kinetic characterization of the chemotactic protein from *Escherichia coli*, CheY. Kinetic analysis of the inverse hydrophobic effect. *Biochemistry*, **33**, 5858–5866.
62. Strop, P., Marinescu, A. M. & Mayo, S. L. (2000). Structure of a protein G helix variant suggests the importance of helix propensity and helix dipole interactions in protein design. *Protein Sci.* **9**, 1391–1394.
63. Radzicka, A. & Wolfenden, R. (1988). Comparing the polarities of the amino acids: side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution. *Biochemistry*, **27**, 1664–1670.
64. Zhang, C., Vasmatzis, G., Cornette, J. L. & DeLisi, C. (1997). Determination of atomic desolvation energies from the structures of crystallized proteins. *J. Mol. Biol.* **267**, 707–726.
65. Fauchere, J. & Pliska, V. (1983). Hydrophobic parameters of amino acid side-chains from the partitioning of *N*-acetyl-amino-acid amides. *Eur. J. Med. Chem.* **18**, 369–375.
66. Nozaki, Y. & Tanford, C. (1971). The solubility of amino acids and two glycine peptides in aqueous ethanol and dioxane solutions. Establishment of a hydrophobicity scale. *J. Biol. Chem.* **246**, 2211–2217.
67. Levitt, M. (1976). A simplified representation of protein conformations for rapid simulation of protein folding. *J. Mol. Biol.* **104**, 59–107.

68. Roseman, M. A. (1988). Hydrophilicity of polar amino acid side-chains is markedly reduced by flanking peptide bonds. *J. Mol. Biol.* **200**, 513–522.
69. Shirley, B. A., Stanssens, P., Hahn, U. & Pace, C. N. (1992). Contribution of hydrogen bonding to the conformational stability of ribonuclease T1. *Biochemistry*, **31**, 725–732.
70. Chen, Y. W., Fersht, A. R. & Henrick, K. (1993). Contribution of buried hydrogen bonds to protein stability. The crystal structures of two barnase mutants. *J. Mol. Biol.* **234**, 1158–1170.
71. Chakrabartty, A., Kortemme, T. & Baldwin, R. L. (1994). Helix propensities of the amino acids measured in alanine-based peptides without helix-stabilizing side-chain interactions. *Protein Sci.* **3**, 843–852.
72. Gilis, D. & Rooman, M. (2000). PoPMuSiC, an algorithm for predicting protein mutant stability changes: application to prion proteins. *Protein Eng.* **13**, 849–856.
73. Lazaridis, T. & Karplus, M. (1999). Discrimination of the native from misfolded protein models with an energy function including implicit solvation. *J. Mol. Biol.* **288**, 477–487.
74. Gatchell, D. W., Dennis, S. & Vajda, S. (2000). Discrimination of near-native protein structures from misfolded models by empirical free energy functions. *Proteins: Struct. Funct. Genet.* **41**, 518–534.
75. Motoc, I. & Marshall, G. R. (1985). Van der Waals volume fragmental constants. *Chem. Phys. Letters*, **116**, 415–419.
76. Creighton, T. E. (1993). Proteins: structures and molecular properties. In *Proteins: Structures and Molecular Properties* (Creighton, T. E., ed.), 2nd edit., pp. 153–155, W.H. Freeman, New York.
77. Ippolito, J. A., Alexander, R. S. & Christianson, D. W. (1990). Hydrogen bond stereochemistry in protein structure and function. *J. Mol. Biol.* **215**, 457–471.
78. Hurley, J. H., Mason, D. A. & Matthews, B. W. (1992). Flexible-geometry conformational energy maps for the amino acid residue preceding a proline. *Biopolymers*, **32**, 1443–1446.
79. Grantcharova, V. P., Riddle, D. S. & Baker, D. (2000). Long-range order in the src SH3 folding transition state. *Proc. Natl Acad. Sci. USA*, **97**, 7084–7089.
80. Covalt, J. C., Jr, Roy, M. & Jennings, P. A. (2001). Core and surface mutations affect folding kinetics, stability and cooperativity in IL-1 beta: does alteration in buried water play a role? *J. Mol. Biol.* **307**, 657–669.
81. Pitt, W. R. & Goodfellow, J. M. (1991). Modelling of solvent positions around polar groups in proteins. *Protein Eng.* **4**, 531–537.
82. Roe, S. M. & Teeter, M. M. (1993). Patterns for prediction of hydration around polar residues in proteins. *J. Mol. Biol.* **229**, 419–427.
83. Petukhov, M., Cregut, D., Soares, C. M. & Serrano, L. (1999). Local water bridges and protein conformational stability. *Protein Sci.* **8**, 1982–1989.
84. Harpaz, Y., Elmasry, N., Fersht, A. R. & Henrick, K. (1994). Direct observation of better hydration at the N terminus of an alpha-helix with glycine rather than alanine as the N-cap residue. *Proc. Natl Acad. Sci. USA*, **91**, 311–315.
85. Pisabarro, M. T. & Serrano, L. (1996). Rational design of specific high-affinity peptide ligands for the Abl-SH3 domain. *Biochemistry*, **35**, 10634–10640.
86. Mateu, M. G. & Fersht, A. R. (1998). Nine hydrophobic side chains are key determinants of the thermodynamic stability and oligomerization status of tumour suppressor p53 tetramerization domain. *EMBO J.* **17**, 2748–2758.
87. Wang, Y., Shen, B. J. & Sebald, W. (1997). A mixed-charge pair in human interleukin 4 dominates high-affinity interaction with the receptor alpha chain. *Proc. Natl Acad. Sci. USA*, **94**, 1657–1662.
88. Hage, T., Sebald, W. & Reinemer, P. (1999). Crystal structure of the interleukin-4/receptor alpha chain complex reveals a mosaic binding interface. *Cell*, **97**, 271–281.
89. Chinea, G., Padron, G., Hooft, R. W., Sander, C. & Vriend, G. (1995). The use of position-specific rotamers in model building by homology. *Proteins: Struct. Funct. Genet.* **23**, 415–421.
90. Hooft, R. W., Sander, C. & Vriend, G. (1996). Positioning hydrogen atoms by optimizing hydrogen-bond networks in protein structures. *Proteins: Struct. Funct. Genet.* **26**, 363–376.
91. Nielsen, J. E., Andersen, K. V., Honig, B., Hooft, R. W., Klebe, G., Vriend, G. & Wade, R. C. (1999). Improving macromolecular electrostatics calculations. *Protein Eng.* **12**, 657–662.
92. Wray, J. W., Baase, W. A., Lindstrom, J. D., Weaver, L. H., Poteete, A. R. & Matthews, B. W. (1999). Structural analysis of a non-contiguous second-site revertant in T4 lysozyme shows that increasing the rigidity of a protein can enhance its stability. *J. Mol. Biol.* **292**, 1111–1120.
93. Takano, K., Yamagata, Y., Funahashi, J., Hioki, Y., Kuramitsu, S. & Yutani, K. (1999). Contribution of intra- and intermolecular hydrogen bonds to the conformational stability of human lysozyme. *Biochemistry*, **38**, 12698–12708.

Edited by A. R. Fersht

(Received 21 February 2002; received in revised form 21 February 2002; accepted 3 May 2002)

http://www.academicpress.com/jmb

Supplementary Material comprising six Tables and one Figure is available on IDEAL