J. Mol. Biol. (1996) **257**, 1112–1126

JMB

Stability Changes upon Mutation of Solvent-accessible Residues in Proteins Evaluated by Database-derived Potentials

Dimitri Gilis* and Marianne Rooman

UCMB, Université Libre de Bruxelles, CP160/16 a.v. F. Roosevelt 1050 Brussels Belgium

**Corresponding author*

The stability changes in peptides and proteins caused by the substitution of a single amino acid, which can be measured experimentally by the change in folding free energy, are evaluated here using effective potentials derived from known protein structures. The analysis is focused on mutations of residues that are accessible to the solvent. These represent in total 106 mutations, introduced at different sites in barnase, bacteriophage T4 lysozyme and chymotrypsin inhibitor 2, and in a synthetic helical peptide. Assuming that the mutations do not modify the backbone structure, the changes in folding free energies are computed using various types of database-derived potentials and are compared with the measured ones. Distance-dependent residue–residue potentials are found to be inadequate for estimating the stability changes caused by these mutations, as they are dominated by hydrophobic interactions, which do not play an essential role at the protein surface. On the contrary, the potentials based on backbone torsion angle propensities yield quite good results. Indeed, for a subset of 96 out of the 106 mutations, the computed and measured changes in folding free energy correlate with a linear correlation coefficient of 0.87. Moreover, the ten mutations that are excluded from the correlation either seem to cause modifications of the backbone structure or to involve strong hydrophobic interactions, which are atypical for solvent-accessible residues. We find furthermore that raising the ionic strength of the solvent used for measuring the changes in folding free energies improves the correlation, as it tends to mask the electrostatic interactions. When adding to these 106 mutations 44 mutations performed in staphylococcal nuclease and chemotactic protein, which were first discarded because some of them were suspected to affect the backbone conformation or the denatured state, the correlation between measured and computed folding free energy changes remains quite good: the correlation coefficient is 0.86 for 135 out of the 150 mutations. The success of the backbone torsion potentials in predicting stability changes indicates that the approximations made for deriving these potentials are adequate. It suggests moreover that the local interactions along the chain dominate at the protein surface.

Keywords: single-site mutations; folding free energies; backbone torsion potentials; residue–residue interaction potentials; protein stability

Introduction

The principles that rule protein stability have begun to be better understood since protein engineering experiments such as site-directed mutagenesis have provided experimental data on the relative stability of a lot of mutants. The common procedure consists of substituting one or several residues and determining the corresponding stability changes by measuring the changes in the free energy of unfolding upon denaturation (O'Neil & DeGrado, 1990; Serrano *et al*., 1990, 1992a,b; Matoushek *et al*., 1989; Sali *et al*., 1991; Horovitz *et al*., 1992; Dao-Pin *et al*., 1990; Hu *et al*., 1992; Jackson & Fersht, 1994; Blaber *et al*., 1993; Zhang *et al*., 1995; Itzhaki *et al*., 1995; Shortle *et al*., 1990; Green *et al*., 1992; Muñoz *et al*., 1994; López-Hernández & Serrano, 1995). These data are then interpreted in terms of the modification of the interactions that stabilize the tertiary structures (for

0022–2836/96/151112–15 $18.00/0

a review, see Fersht & Serrano, 1993). This strategy has confirmed the dominating influence of the hydrophobic interactions in the protein core, but has also revealed the non-negligible role of other interactions, such as hydrogen bond and electrostatic interactions, which must all be satisfied to achieve protein stability. However, the specific amount of stabilization provided by the different components is not known with precision, because it is highly contextual and dependent on the spatial environment in the protein.

In parallel with the increasing body of experimental results, the problem of protein stability has been tackled from a theoretical point of view, with the ultimate aim of being capable of designing protein sequences with a pre-defined structure. More realistically, most theoretical approaches focus on single-site mutations that are assumed to be sufficiently neutral to keep the backbone conformation almost unperturbed, and estimate the corresponding stability change on the basis of various energy criteria.

The first procedures aimed at predicting the stability changes of mutant proteins were free energy calculations with detailed atomic models coupled to semi-empirical force fields (Basch *et al*., 1987; Dang *et al*., 1989; Tidor & Karplus, 1991). But these procedures are so computer-time-consuming that they cannot at present be applied to a large number of mutations. This has prompted the development of faster methods based on rougher descriptions of protein structures. One of these methods uses a simplified force field combined with a search in a limited conformational space (Lee & Levitt, 1991; Lee, 1994). Methods based on even more simplified models estimate folding free energies using effective potentials derived from known protein structures, in particular hydrophobic potentials (Koehl & Delarue, 1994), secondary structure potentials (Muñoz & Serrano, 1994), residue contact potentials (Miyazawa & Jernigan, 1994) and distance-dependent residue–residue interaction potentials (Sippl, 1995). Still others relate the stability changes to the shape, flexibility and volume of the substituted amino acids (van Gunsteren & Mark, 1992), to the number of methylene and methyl groups in the environment of the mutated residues (Serrano *et al*., 1992b), to the number of surrounding α-carbon atoms (Shortle *et al*., 1990), or to the cavity formation in the protein interior resulting from mutating a large into a smaller amino acid (Eriksson *et al*., 1992; Kocher, J. P., Prévost, M., Lee, B. K. & Wodak, S. J., unpublished data). Most of these methods (Dang *et al*., 1989; Tidor & Karplus, 1991; Lee & Levitt, 1991; Lee, 1994; Koehl & Delarue, 1994; Miyazawa & Jernigan, 1994; van Gunsteren & Mark, 1992; Eriksson *et al*., 1992; Kocher *et al*., unpublished data) have been applied to mutations of residues that are buried in the protein core, where the hydrophobic effect is predominant. This choice is quite natural for procedures using hydrophobicity potentials or residue pair potentials, in which the hydrophobicity component is known to dominate (Casari & Sippl, 1992; Bryant & Lawrence, 1993; Kocher *et al*., 1994), or when focusing on cavity formation or on the volume and shape of substituted amino acids.

The computed stability changes are usually compared with the experimentally measured changes in folding free energies (Serrano *et al*., 1992b; Dang *et al*., 1989; Tidor & Karplus, 1991; Lee, 1994; Koehl & Delarue, 1994; Muñoz & Serrano, 1994; Miyazawa & Jernigan, 1994; Sippl, 1995; Shortle *et al*., 1990; Eriksson *et al*., 1992; Kocher *et al*., unpublished data), but sometimes they are related to the measured activity changes (Lee & Levitt, 1991; van Gunsteren & Mark, 1992), though stability and activity do not always go together (Shoichet *et al*., 1995). The reported correlations between computed and measured quantities are all reasonably good. However, they are always restricted to selected mutations in a single protein, and usually even at a single site. As soon as mutations in different sites and different proteins are mixed, the correlations seem to break down. These effects are usually attributed to shortcomings of the energy criteria used or to the various approximations made in deriving them. It has tentatively been attributed to the fact that the unfolded state could be different in the different sites and different proteins (Miyazawa & Jernigan, 1994).

Here, we consider single-site mutations of residues that are solvent-accessible and compute the change in folding free energy using different types of database-derived potentials, assuming that the backbone conformation remains unchanged upon mutation. The computed folding free energy changes are compared with the experimentally measured ones. We start by presenting the ensemble of mutations that we consider, and continue by exposing the different formalisms that have been developed for deriving potentials from protein structure data. This issue is quite important. Indeed, the different formalisms do not agree on the interpretation of the computed quantities, and hence on the applicability of the database-derived potentials for computing folding free energies. The choice of the correct formalism constitutes the basis of our whole approach.

Ensemble of Mutations Considered

The present analysis is restricted to mutations of residues situated at the protein surface, which expose at least 50% of their accessible surface area to the solvent as measured by the algorithm SurVol (Alard, 1991). In addition, only mutants with a single amino acid substitution are considered. The ensemble of mutations is given in Table 1A and B. It comprises 38 mutations in barnase at eight different sites (Serrano *et al*., 1990, 1992a,b; Matoushek *et al*., 1989; Sali *et al*., 1991; Horovitz *et al*., 1992), 32 mutations in T4 phage lysozyme at six sites (Dao-Pin *et al*., 1990; Hu *et al*., 1992; Blaber *et al*., 1993; Zhang *et al*., 1995), 17 mutations in

Table 1. Set of considered mutations

Protein / peptide[a]	Sequence position	Mutated residues	Introduced mutations	(φ, ψ, ω) Domain[b]	Secondary structure[c]	Accessibility[d] (%)
A. *Set of 106 mutations in barnase, T4 lysozyme, chymotrypsin inhibitor 2 and the ODG peptide*						
Barnase	8	Asp	Ala, Gly, Ser	A	H	77.6
	12	Asp	Ala, Gly, Ser	A	H	52.9
	16	Thr	Ala, Arg, Gly, Ser	A	H	57.8
	28	Ser	Ala, Gly, Glu	A	H	60.7
	31	Gln	Ala, Gly, Ser	A	H	62.4
	32	Ala	Gly, Lys, Arg, Met, Leu, Ser, Gln, Glu, Asn, Phe, Asp, His, Thr, Ile, Tyr, Val, Trp, Cys, Pro	A	H	76.3
	55	Ile	Ala, Val	P	E	54.9
	105	Thr	Val	B	C	62.9
T4 lysozyme	44	Ser	Ala, Cys, Asp, Glu, Phe, His, Ile, Lys, Leu, Met, Asn, Pro, Gln, Arg, Gly, Thr, Val, Trp, Tyr	A	H	54.1
	82	Ala	Pro	A	T	78.3
	93	Ala	Pro	A	H	67.2
	113	Gly	Ala	A	T	62.6
	131	Val	Ala, Thr, Leu, Met, Ile, Glu Ser, Asp, Gly	A	H	66.3
	144	Asn	Asp	A	H	66.8
Chymotrypsin inhibitor 2	33	Glu	Gln, Asp, Asn	A	H	53.7
	34	Glu	Gln, Asp, Asn	A	H	51.5
	37	Lys	Ala, Gly	A	H	61.1
	41	Gln	Ala, Gly	A	H	76.0
	44	Pro	Ala	C	T	72.5
	45	Glu	Ala	C	T	50.3
	52	Pro	Ala	P	E	54.4
	56	Ile	Ala	P	C	94.2
	58	Thr	Ala, Asp	P	C	69.9
	72	Lys	Asn	C	T	93.9
ODG peptide	14	Gly	Ala, Cys, Asp, Glu, Phe, His, Ile, Lys, Leu, Met, Asn, Pro, Gln, Arg, Ser, Thr, Val, Trp, Tyr	A	H	Fully accessible
B. *Set of 44 mutations in staphylococcal nuclease and chemotactic protein, some of which were suspected to modify the backbone conformation or the denatured state*						
Staphyloccal nuclease	11	Pro	Gly, Ala	P	E	74.1
	13	Thr	Gly, Ala	B	E	60.7
	29	Gly	Ala, Val	G	T	85.4
	30	Gln	Gly, Ala	B	E	54.2
	31	Pro	Gly, Ala	P	E	53.5
	47	Pro	Gly, Ala	A	T	82.4
	51	Val	Gly, Ala	B	S	60.6
	60	Ala	Gly, Val	A	H	51.0
	68	Asn	Gly, Ala	C	T	67.0
	80	Gln	Gly, Ala	P	C	60.5
	82	Thr	Gly, Ala	B	B	50.1
	85	Tyr	Gly, Ala	C	T	70.4
	86	Gly	Ala, Val	G	S	50.2
	96	Gly	Ala, Val	G	T	61.9
	113	Tyr	Gly, Ala	P	C	74.1
	115	Tyr	Gly, Ala	B	B	67.6
	123	Gln	Gly, Ala	A	H	58.8
	141	Ser	Gly, Ala	X	C	83.2
Chemotactic protein	14	Phe	Asn	B	C	62.1
	48	Ala	Gly	C	T	80.6
	74	Ala	Gly	C	T	78.6
	77	Ala	Gly	C	T	81.8
	80	Ala	Gly	C	T	77.3
	88	Ala	Gly	C	S	61.7
	90	Ala	Gly	P	C	50.7
	114	Ala	Gly	A	H	62.1

[a] Proteins or peptides in which the mutations are introduced. The PDB code (Bernstein *et al*., 1977) of the barnase structure used here is lRNB, that of T4 lysozyme is 1LYD, that of chymotrypsin inhibitor 2 is 2CI2, that of staphylococcal nuclease is 1STN and that of the chemotactic protein is 2CHF. ODG peptide refers to the synthetic α-helical ODG peptide of O'Neil & DeGrado (1990).

[b] The (φ, ψ, ω) domains of the mutated residues in the wild-type structure, using the definitions of Rooman *et al*. (1991, 1992) and Kocher *et al*. (1994). A corresponds to right-handed α-helical conformations, C to 3_{10}-helix, G to left-handed helix, B and P to extended conformations, with B comprising more particularly β-strands and P poly(proline) type conformations, and X denotes residues for which one of the torsion angles could not be determined.

[c] Secondary structure of the mutated residue calculated by DSSP (Kabsch & Sander, 1983). H means α-helix, C random coil, E β-strand, B isolated β, S bend and T turn.

[d] Solvent accessibility (in %), defined as the solvent-accessible surface area of the residue in its parent protein, computed by SurVol (Alard, 1991), divided by the solvent-accessible surface area of the residue in an extended tripeptide Gly-X-Gly conformation (Rose *et al*., 1985). As the precise structure of the ODG peptide has not been determined (it is only known to be helical) the exact accessibility of the residue Gly14 cannot be computed.

chymotrypsin inhibitor 2 at ten sites (Jackson & Fersht, 1994; Itzhaki *et al*., 1995), 19 mutations in a synthetic α-helical peptide of 29 residues (O'Neil & DeGrado, 1990), which will be referred to as ODG peptide, 36 mutations in staphylococcal nuclease at 18 sites (Shortle *et al*., 1990; Green *et al*., 1992) and eight mutations in chemotactic protein at eight sites (Muñoz *et al*., 1994; López-Hernández & Serrano, 1995). When the folding free energy difference of the same mutation is measured at different pH values or salt concentrations, we take the value measured at the highest pH or highest salt concentration, except in the section where we analyse the influence of the experimental conditions.

In total, we consider 150 mutations of residues that are solvent-accessible, introduced in six different proteins or peptides, at different sites, and in different secondary structure with, however, a majority of helices (see Table 1A and B). This corresponds to an attempt at collecting all single-site mutations, performed on surface residues, whose folding free energies have been measured experimentally. The only mutations of solvent-accessible residues that we discard (to our knowledge) are those in Protein G (Minor & Kim, 1994), because the folding free energy changes are measured with respect to a pseudo wild-type with three substituted residues, whose tertiary structure has not been determined.

The main part of our analysis will be performed on the 106 mutations in barnase, T4 phage lysozyme, chymotrypsin inhibitor 2 and ODG peptide (Table 1A). The 44 mutations in staphylococcal nuclease and chemotactic protein (Table 1B) will be discussed separately, because, according to Shortle *et al*. (1990), Green *et al*. (1992) and Muñoz *et al*. (1994), at least a subset of these mutations seems to affect the backbone structure or even the denatured state, which contradicts the basic assumptions of our approach (see Methods).

Formalisms for Deriving Effective Potentials from Known Structures

The most widely used approach for deriving effective potentials from an ensemble of experimentally determined protein structures consists of computing frequencies of sequence and structure features, and converting these frequencies into free energies (Sippl, 1990, 1993, 1995; Wodak & Rooman, 1993; Rooman & Wodak, 1995). In concrete terms, the sequences S are divided into sequence elements s (e.g. residues or residue pairs), and the conformations C into structural states c (e.g. ranges of torsion angles, inter-residue distances or solvent-accessible surface areas). The frequencies of c and s in the protein dataset are computed, yielding an estimation of the probability of c, $P(c)$, and of the conditional probability of c knowing s, $P(c|s)$. Essentially two formalisms have been developed to convert these probabilities into free energies. In terms of frequencies, the same quantity is computed in both formalisms, but it is interpreted differently in terms of free energy.

In the first formalism (Sippl, 1990, 1993; Wodak & Rooman, 1993), the probabilities $P(c|s)$ and $P(c)$ are converted into the free energy $\mathcal{G}^S(C)$ by assuming that the structural states c follow a Boltzmann-type distribution. This yields:

$$\mathcal{G}^S(C) - \mathcal{G}(C) + kT \log \frac{\mathcal{Z}^S}{\mathcal{Z}} = -kT \sum_{i,j} \log \frac{P(c_i|s_j)}{P(c_i)} \quad (1)$$

where k is the Boltzmann's constant and T a conformational temperature (Pohl, 1971), taken to be room temperature. The indices i and j indicate the position(s) along the sequence of the structural states and sequence elements, respectively. $\mathcal{Z}^S$ is the partition function of the system. $\mathcal{Z}$ and $\mathcal{G}(C)$ are the partition function and the free energy of a non-specific sequence. Thus, the computed quantity, which appears on the right-hand side of equation (1), does not correspond to the free energy $\mathcal{G}^S(C)$, but to the difference $\mathcal{G}^S(C) - \mathcal{G}(C)$ plus a term containing partition functions. Because of the dependence of the latter term on the sequence, this equation can in principle only be applied to compare different structures of the same sequence. It is, in particular, not suited to compute the energy difference of two mutant proteins, unless the additional approximation is made that the mutation does not modify the partition function.

In the second derivation (Rooman & Wodak, 1995), it is argued that the quantity $\mathcal{G}^S(C)$ does not approximate the true free energy of the ensemble of conformations C of the system S, because it is computed without correcting for the many-body effect that arises from the presence of other residues and screens out the correlations between s and c. The need for correcting for this effect is most clearly seen when the structural states c are inter-residue distance ranges. Indeed, the most populated states c, and hence the conformations C with lowest value of $\mathcal{G}^S(C)$, are those in which residues are not in contact, whereas the conformations with lowest free energy usually have many inter-residue contacts. Note that the quantity $\mathcal{G}^S(C) - \mathcal{G}(C)$ includes a correction for the many-body effects, but it has no clear physical meaning. When these effects are properly taken into account, it was shown by Rooman & Wodak (1995) that the partition function term drops out of the equations. The probability ratios that appear in the right-hand side of equation (1) turn out to approximate the difference between the true free energy $G^S(C)$ and the free energy G^S of a denatured-like state of S in which the conformational states c and the sequence elements s are uncorrelated:

$$\Delta G^S(C) = G^S(C) - G^S = -kT \sum_{i,j} \log \frac{P(c_i|s_j)}{P(c_i)} \quad (2)$$

This free energy difference, $\Delta G^S(C)$, is referred to as the folding free energy. It can be used for comparing different structures of the same sequence, as well as

different sequences with the same structure. However, the remaining question is whether the considered denatured-like state is a good approximation of the true one, or if other approximations must be used. Preliminary results were positive in that respect, but not conclusive (Rooman & Wodak, 1995).

There is in fact another, earlier, derivation of database-derived potentials (Miyazawa & Jernigan, 1985, 1994), which is specifically designed to compute residue–residue contact potentials. It takes explicitly solvent molecules into account, and uses Boltzmann law, *via* Bethe approximation, to convert frequencies into energies. It uses a correction for the many-body effects that is more similar to the correction performed in the first (eqn (1)) than in the second derivation (eqn (2)). The computed quantity corresponds to an energy difference, between the formation of residue–residue *versus* residue-solvent contacts, and hence does not include a partition function term. This derivation is confronted with the problem of estimating the number of contacts between solvent molecules, and, when applied to compute folding free energies of mutant proteins, of modelling the denatured state (Miyazawa & Jernigan, 1994).

In this paper, the validity of the second formalism is stated and, as a consequence, the applicability of database-derived potentials for computing stability changes upon mutation. The folding free energies ΔG are thus computed using equation (2). Two main types of potentials are considered, the backbone torsion potentials, which are derived from frequencies of backbone torsion angle domains, and the residue–residue potentials, which are based on the propensities of residue pairs to be separated by a given spatial distance. The details of these potentials, of the computation of the changes in folding free energy $\Delta\Delta G$ upon mutation and their comparison with the measured $\Delta\Delta G$ values are given in Methods.

Correlation Between Measured and Computed $\Delta\Delta G$ Values Using Backbone Torsion Potentials

The changes in folding free energy $\Delta\Delta G$ between wild-type and mutant proteins are computed using backbone torsion potentials for the 106 mutations of solvent-accessible residues listed in Table 1A, and are compared with the experimentally measured $\Delta\Delta G$ values. Four variants of backbone torsion potentials are used, the short and middle range torsion → residue and residue → torsion potentials (see Methods). The $\Delta\Delta G$ values are depicted in Figure 1(a) and (b) for the two torsion → residue potentials; the two residue → torsion potentials give similar, but slightly less good, results and are hence not shown.

For the ensemble of 106 mutants, the computed and measured $\Delta\Delta G$ values are only weakly correlated, with linear correlation coefficients between 0.60 and 0.67 for the different torsion potentials. However, when visualizing the $\Delta\Delta G$ values (Figure 1), it appears that the weak correlation is in fact due to a few mutations, which are clearly apart from the main group. To identify these outsiders objectively, we use an automatic sorting procedure that rejects one mutation at a time from the original ensemble, until the correlation coefficient exceeds a certain value. The rejected mutation is the mutation that, when discarded from the ensemble, gives rise to the highest correlation coefficient for the remaining mutations.

With this sorting procedure, we find that dropping ten out of the 106 mutations increases the correlation coefficient to 0.87 for both the short and middle range torsion → residue potentials, and to 0.82 and 0.86 for the short and middle range residue → torsion potentials, respectively. The two torsion → residue potentials thus yield the highest scores, in agreement with previous findings (Kocher *et al*., 1994).

Strikingly, nine out of the ten mutations that are rejected first by the sorting procedure are the same for all the torsion potentials. These mutations are: Thr16 → Ser, Thr16 → Gly, Thr105 → Val and Ala32 → Pro in barnase; Ser44 → Pro, Ala82 → Pro and Ala93 → Pro in T4 lysozyme; Pro52 → Ala in chymotrypsin inhibitor 2; and Gly14 → Pro in the ODG peptide. The tenth rejected mutation is Asp8 → Ala in barnase for the two short range potentials, and Pro44 → Ala in chymotrypsin inhibitor 2 for the two middle range potentials.

The fact that the folding free energy of the mutation Asp8 → Ala in barnase, which involves the loss of electrostatic interactions, is better estimated by middle range potentials, and that the mutation Pro44 → Ala in chymotrypsin inhibitor 2, which is situated in an αβ turn, is better described by the short range potentials, can be taken to mean that short and middle range potentials represent somewhat different interactions, which are each better suited to different applications. This interpretation is supported by the fact that the short range potentials were found to be superior on the basis of the ability to recognize native sequence–structure matches from a set of alternatives (Kocher *et al*., 1994), whereas the opposite trend appears in the present analysis. This interpretation can seem surprising *a priori*, as the middle range potentials, which combine all the residue influences in the $[i-8,\ i+8]$ sequence window, include the short range ones, which are restricted to the $[i-1,\ i+1]$ window. However, the specific interactions in the $[i-1,\ i+1]$ window have less weight in the middle than in the short range potentials, because they are averaged with many more interactions.

Among the 106 mutations considered, 94 are performed in helices, three in β-strands, five in turns and four in coil regions (Table 1A). In spite of a large majority of helices, the mutations introduced in different secondary structures are found to be roughly at the same distance from the regression line, on average. This result is true for the four types

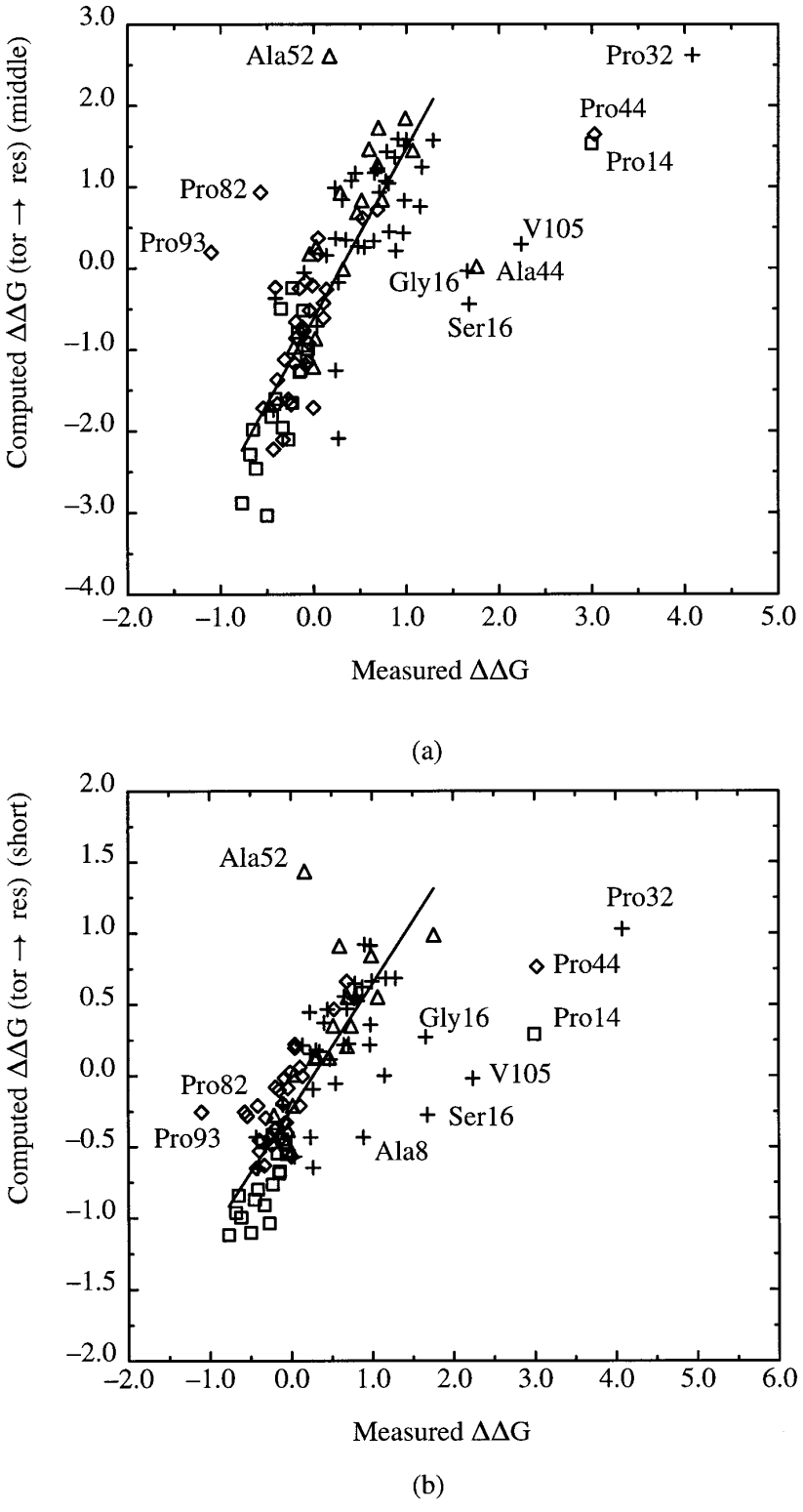

Figure 1. ΔΔ*G* values computed from backbone torsion potentials as a function of the measured ΔΔ*G* values, for the 106 mutants listed in Table 1A. The mutations in barnase, T4 lysozyme, chymotrypsin inhibitor 2 and the ODG peptide are indicated by +, ◇, △ and □, respectively. The measured ΔΔ*G* values are in kcal / mol and are taken from O'Neil & DeGrado (1990), Serrano *et al*. (1990, 1992a,b), Matoushek *et al*. (1989), Sali *et al*. (1991), Horovitz *et al*. (1992), Dao-Pin *et al*. (1990), Hu *et al*. (1992), Blaber *et al*. (1993), Jackson & Fersht (1994), Itzhaki *et al*. (1995) and Zhang *et al*. (1995). The computed ΔΔ*G* values are formally also in kcal / mol. The plotted lines correspond to the regression lines obtained with ten out of the 106 mutations excluded by our sorting procedure. The excluded mutations are indicated by the name of the mutant amino acid followed by its position in the sequence. (a) Computed ΔΔ*G* values obtained with the middle range torsion → residue potential. The linear correlation coefficient between measured and computed ΔΔ*G* values is equal to 0.67 for all 106 mutations, and to 0.87 when the ten mutations are rejected. The probability $\mathscr{P}$ that the latter correlation is obtained by chance is $\mathscr{P} = 0.000000$. The equation of the regression line is: $y = 2.10x - 0.62$. (b) Computed ΔΔ*G* values obtained with the short range torsion → residue potential. The linear correlation coefficient between measured and computed ΔΔ*G* values is equal to 0.67 for all 106 mutations, and to 0.87 when the ten mutations are rejected ($\mathscr{P} = 0.000000$). The equation of the regression line is: $y = 0.88x - 0.24$.

of backbone torsion potentials. This indicates that these potentials can be used to evaluate folding free energy changes, irrespective of the secondary structures in which the mutations are performed.

Thus, the stability changes caused by 96 out of the 106 mutations are reliably quantified by the backbone torsion potentials. The ten remaining mutations are therefore expected to have unusual characteristics, which we analyse in the next section.

Why do certain mutations depart from the regression line?

Given that among the ten mutations that are responsible for the low overall correlation

coefficients, only nine coincide for the short and middle range potentials, a total of 11 mutations must be analysed. Among these 11 mutations, seven involve proline: Ala32 → Pro in barnase, Ser44 → Pro, Ala82 → Pro and Ala93 → Pro in T4 lysozyme; Pro44 → Ala and Pro52 → Ala in chymotrypsin inhibitor 2; and Gly14 → Pro in the ODG peptide. Moreover, these are the only considered mutations that involve proline.

The reason why the computed $\Delta\Delta G$ values for these Pro mutations are different from the measured ones is easily understood, considering that our approach rests on the assumption that the backbone conformation remains unchanged upon mutation. Indeed, the mutations Ala32 → Pro in barnase, Ser44 → Pro in T4 lysozyme and Gly14 → Pro in the ODG peptide introduce a Pro in the middle of a helix or near its C terminus. This can be expected to induce a kink in the helix (Piela *et al*., 1987), thereby modifying the backbone conformation. Our $\Delta\Delta G$ computation indicates that these mutations are destabilizing, but in fact they are even more destabilizing than what is computed, probably because the kink in the helix destroys other interactions in the vicinity.

Similarly, the mutation Ala93 → Pro in T4 lysozyme introduces a proline at the first position in the helix. Since Pro is quite unfavorable at that position, but quite favorable as N-cap (Richardson & Richardson, 1988)†, this mutation probably shortens the helix by one residue, and thus also modifies the backbone conformation. The same situation occurs with the Ala82 → Pro mutation in T4 lysozyme. Indeed, though DSSP (Kabsch & Sander, 1983) assigns this residue as being in a turn between two helices on the basis of the hydrogen bonding pattern, it can be considered as the first position of the helix on the basis of the (ϕ, ψ)-angles, with the N-terminal helix turn being somewhat looser. According to our computation, these mutations are destabilizing, though they are stabilizing in reality. However, if the folding free energy of the mutants were computed in their true structure, in which the considered helix starts one residue further on, they would probably be found to be stabilizing‡. It must be noted that the departure of the two mutations Ala82 → Pro and Ala93 → Pro in T4 lysozyme from the regression line could also be attributed, at least in part, to the large experimental error in the measured $\Delta\Delta G$ values. These errors are indeed equal to ± 0.4 kcal/mol, whereas they are ± 0.1 kcal/mol at the most for the other considered mutations.

The last two mutations that involve proline, also seem to modify the backbone structure. The Pro52 → Ala mutation in chymotrypsin inhibitor 2 is introduced at the last position of a β-strand. This residue has its backbone torsion angles in the P domain, corresponding to an extended conformation often adopted by proline, but seldom by alanine. This explains why this mutation is computed to be highly destabilizing. In reality, it is only slightly destabilizing, probably because the Ala52 in the mutant structure falls in another torsion angle domain. Finally, the mutation Pro44 → Ala in chymotrypsin inhibitor 2 occurs at the second (L2) position of an αBAAβ turn (Wintjens *et al*., 1996). This mutation is very destabilizing and is not computed as destabilizing enough, especially by the middle range potentials. This turn is a recurrent motif of which five examples have been found in a representative protein dataset (unpublished results). Strikingly, four out of these five examples have Pro at position L2. This residue thus seems necessary for stabilizing this motif, and mutating it to Ala is expected to modify the type of turn.

These examples show one of the limitations of our method: as soon as the mutation causes a backbone rearrangement that modifies the backbone torsion angle domain of at least one residue, our approach breaks down. It can be envisaged to model these backbone rearrangements, but this is another issue.

The reason why the mutations Thr16 → Ser and Thr16 → Gly in barnase depart from the over-all correlation is different. Though Thr16 is partly exposed to the solvent, with an accessibility of 58%, it packs its methyl group against the aromatic side-chain of Tyr17 to make a hydrophobic interaction. This interaction is lost by mutating Thr into Ser or Gly, which do not contain a methyl group (Serrano *et al*., 1992b). These mutations are quite destabilizing, as measured by $\Delta\Delta G$ values of 1.68 and 1.66 kcal/mol, respectively. The hydrophobic effect is thus very important in this case. Considering that the backbone torsion potentials do not account for this interaction, it is not surprising that the computed $\Delta\Delta G$ values are rather different from the measured ones and predict the mutations to be roughly neutral. Hence, notwithstanding the rather large accessibility of Thr16, the mutations Thr16 → Ser and Thr16 → Gly would fit better in the group of buried mutations, which we have not analysed here.

As for the mutation Asp8 → Ala in barnase, whose destabilizing effect is underestimated especially by the short range potentials, it involves the loss of electrostatic interactions, which are not explicitly taken into account in the torsion potentials. Note that other mutations, such as Asp12 → Ala in barnase, have the same property but are closer to the regression line and are not rejected by our procedure. The reason for this difference is that these mutations, and in particular Asp12 → Ala, are measured at higher salt concentrations, which tend to mask the electrostatic interactions, as will be discussed in the next section.

† Note that the helix assignment used in the present work is that of DSSP (Kabsch & Sander, 1983), which differs from the assignment of Richardson & Richardson (1988), the latter containing one residue more at both helix termini. The helix N-cap considered here thus corresponds to the first position in the helix described by Richardson & Richardson (1988).

‡ See Note added in proof.

Finally, the departure of Thr105 → Val in barnase from the over-all correlation is *a priori* less clear. This mutation is highly destabilizing, as measured by a $\Delta\Delta G$ of 2.24 kcal/mol. According to Serrano *et al*. (1992b), this destabilization is explained by the breaking of a hydrogen bond between the side-chain OH group of Thr105 and the $O^{\delta 1}$ of Asp101. In the X-ray structure of barnase that we use (1RNB), this hydrogen bond is not present, but is replaced by a hydrogen bond with a water molecule. Though the breaking of a hydrogen bond is certainly destabilizing, it is surprising that it is so highly destabilizing. Indeed, for hydrogen bonds that have access to water, the usually observed values are in the range 0.0 to 0.5 kcal/mol (Fersht & Serrano, 1993). Another explanation of the large destabilization of this mutation would be that it affects the backbone structure. Thr105 is positioned at the end of a turn between two β-strands. Introducing a Val at that position might cause some structural rearrangements of the type that have been described above for the mutation Pro44 → Ala in chymotrypsin inhibitor and have been observed in turn regions of staphylococcal nuclease (Hynes *et al*., 1994).

Influence of the experimental conditions on the correlation

The $\Delta\Delta G$ values of the 106 mutants studied here are not all measured under the same experimental conditions. In particular, the pH and the ionic strength of the solvent vary, as well as the way in which the proteins are denatured. As these conditions only modify the measured but not the computed $\Delta\Delta G$ values, they can be suspected to affect the correlation between these values.

Urea is used as denaturant for all the mutants of barnase and the ODG peptide, guanidium chloride is used for the chymotrypsin inhibitor 2 mutants and the T4 lysozyme mutants are thermally denatured. Inspection of Figure 1 shows that the lysozyme mutants and the chymotrypsin inhibitor mutants are not further away from the regression line than the other mutants. On the basis of these observations, it can be suggested that the type of denaturation has no effect on the correlation, in agreement with Kellis *et al*. (1989).

To investigate the influence of the ionic strength of the solvent, we dispose of four mutations whose $\Delta\Delta G$ values are measured at different concentrations of NaCl and Na-Mes ([*N*-morpholino]ethanesulphonic sodium): Asp12 → Ala, Thr16 → Arg, Ala32 → Lys and Ser28 → Glu in barnase. The $\Delta\Delta G$ values of these four mutations are situated on both sides of the regression line for the lowest concentration of NaCl. When increasing this concentration, the mutation Thr16 → Arg remains approximately on the regression line, and the three other mutations come closer to it (Figure 2(a)). These four mutations share a common feature: they involve the substitution of a charged by a non-charged residue, or the reverse. This feature explains why better correlations are achieved with higher salt concentrations. The latter tend indeed to mask electrostatic interactions (Serrano *et al*., 1990), which are not explicitly taken into account in the torsion potentials. On the contrary, increasing the Na-Mes concentration does not draw all the mutations closer to the regression line. This suggests that Na-Mes, which is used as part of the buffer in the denaturation experiment, has a different effect from that of NaCl.

We dispose of six mutations whose $\Delta\Delta G$ values are measured at different pH values. These mutations are performed in T4 lysozyme, with pH values between 2 and 5.4. The mutants of barnase and chymotrypsin inhibitor 2 are all measured at pH 6.3 and the mutants of the ODG peptide at pH 7.5. As shown in Figure 2(b), raising the pH decreases the measured $\Delta\Delta G$ for the six lysozyme mutants. For the mutations Gly113 → Ala and Val131 → Ala, this amounts to moving the measured $\Delta\Delta G$ exactly onto the regression line, whereas for the mutations Ala82 → Pro, Ala93 → Pro, Val131 → Thr and Asn144 → Asp the $\Delta\Delta G$ values move further away. This suggests that increasing the pH decreases the measured $\Delta\Delta G$ values, but does not have a systematic effect on the correlation coefficient.

In fact, the two mutations Ala82 → Pro and Ala93 → Pro in T4 lysozyme are rejected by our sorting procedure, and therefore are not included in the set of well-correlating mutations. However, as seen in Figure 2(b), the measured $\Delta\Delta G$ values of these mutations are particularly sensitive to the pH. For these mutations, considering the $\Delta\Delta G$ values measured at low pH would draw them close to the regression line, and would prevent them being rejected by the sorting procedure.

$\Delta\Delta G$ correlation for the mutations in staphylococcal nuclease and chemotactic protein

A total of 44 surface-exposed mutations were considered in these two proteins (Table 1B). They were considered separately from the main group, because some of them were suspected to affect the backbone structure or the denatured state (Shortle *et al*., 1990; Green *et al*., 1992; Muñoz *et al*., 1994), which is in contradiction to the basic assumptions of our approach.

To see where these mutations are positioned relative to the 106 mutations analysed above and, in particular, relative to the subset of 96 mutations whose measured and computed $\Delta\Delta G$ values correlate well, all these mutations are put together in the same graph (Figure 3). Quite surprisingly, we see that most of the 44 additional mutations are rather close to the regression line and that some of them are even on it. In fact, with the short range torsion → residue potential, a correlation coefficient of 0.86 is reached for as many as 135 out of the 150 considered mutations. Thus, only five out of the 44 mutations are rejected, the mutations Gly86 → Ala,

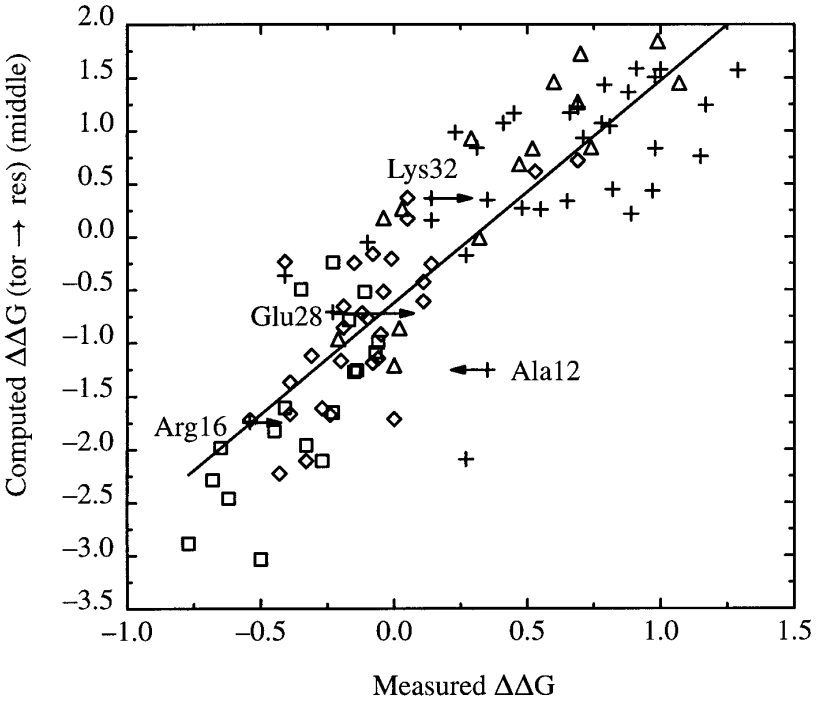

Figure 2. $\Delta\Delta G$ values computed with the middle range torsion → residue potential as a function of the $\Delta\Delta G$ values measured under different experimental conditions, for the 96 mutants that remain after the sorting procedure. See the legend to Figure 1 for details. (a) $\Delta\Delta G$ values measured at different concentrations of NaCl, between 0 mM and 900 mM, taken from Serrano *et al*. (1990) and Sali *et al*. (1991). The four mutations for which such data are available are given by the name of the mutant amino acid followed by its position in the sequence. The different values of $\Delta\Delta G$ measured for these four mutations are indicated by arrows, starting at the lowest NaCl concentration and ending at the highest. (b) $\Delta\Delta G$ values measured at different pH values between 2 and 5.4, which are available for the six mutations for which the mutant amino acid and the sequence position are indicated (Dao-Pin *et al*., 1990; Hu *et al*., 1992; Zhang *et al*., 1995). Two proline residues excluded by the sorting procedure are also shown. The different values of $\Delta\Delta G$ are indicated by arrows, starting at the lowest pH and ending at the highest.

Ala60 → Val, Thr82 → Gly and Gln80 → Gly in staphylococcal nuclease and the mutation Phe14 → Asn in chemotactic protein. Hence, the fraction of rejected mutations here is of the same order as in the main group.

A possible explanation for the departure of some of these mutations from the regression line can be found in the experimental errors in their measured $\Delta\Delta G$ values. These errors are indeed in the range ± 0.1 to 0.5 kcal/mol, whereas for most mutations of the main group they are lower than ± 0.1 kcal/mol. Taking the error bars into account for the mutations that depart most from the regression line, as depicted in Figure 3, shows that the majority of the $\Delta\Delta G$ values could come sufficiently close to the regression line to be considered as well predicted.

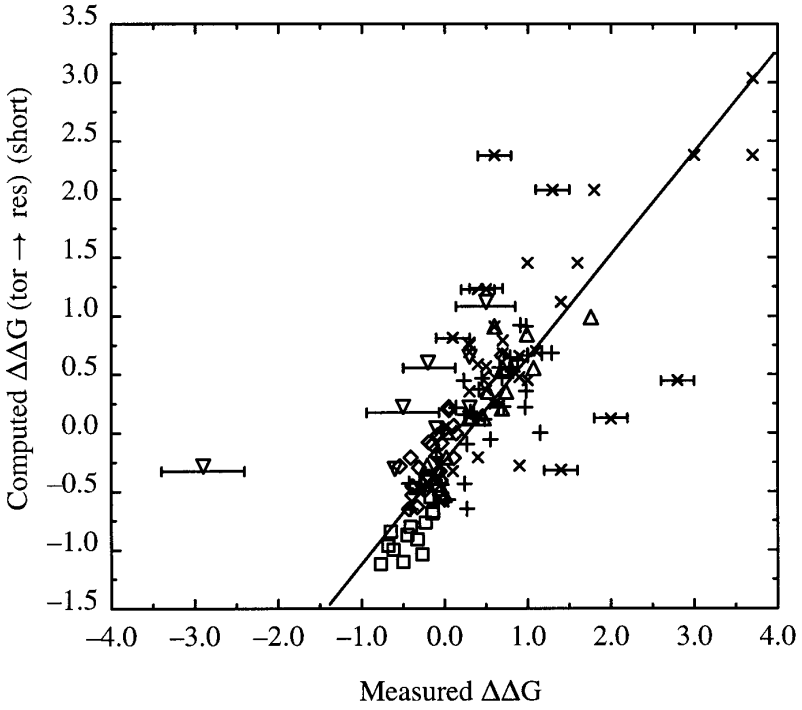

Figure 3. $\Delta\Delta G$ values computed from short range torsion → residue potential as a function of the measured $\Delta\Delta G$ values, for the 150 mutations listed in Table 1. Out of the 44 mutations in Table 1B, 36 are introduced in staphylococcal nuclease (Shortle *et al*., 1990, Green *et al*., 1992) and are symbolized by ×, and eight are introduced in chemotactic protein (Muñoz *et al*., 1994; López-Hernández & Serrano, 1995) and are symbolized by ▽. The experimental error bars of the measured $\Delta\Delta G$ values are depicted for the mutations that depart most from the regression line. The 106 mutations of Table 1A are symbolized as in Figure 1. Only 96 out of these 106 mutations are depicted in this Figure. These are the 96 mutations that remain after the sorting procedure (see the text). The regression line is calculated on these 96 mutations and coincides with the regression line of Figure 1(b).

These results suggest that of the set of 44 mutations in staphylococcal nuclease and chemotactic protein, only some imply significant modifications of the backbone structure or of the denatured state. The mutation that is furthest away from the regression line, and that can by no means be explained by the experimental errors, is the mutation Phe14 → Asn in chemotactic protein. This mutation is the only one that has been analysed in detail experimentally, and for which it has been shown that the major stabilization effect comes from the relative destabilization of the unfolded state and of the kinetic intermediate with respect to the transition state (Muñoz *et al*., 1994). Clearly, the $\Delta\Delta G$ that we compute does not account for this relative destabilization.

Correlation Between Measured and Computed $\Delta\Delta G$ Values Using Residue–Residue Potentials

The C^{μ}–C^{μ} and C^{β}–C^{β} potentials, which describe distance-dependent residue-residue interactions, lead to lower correlation coefficients between computed and measured $\Delta\Delta G$ values than the backbone torsion potentials, as seen in Figure 4. Indeed, the correlations coefficients are of 0.2 and 0.3 for the ensemble of 106 mutants listed in Table 1A, for the C^{μ}–C^{μ} and C^{β}–C^{β} potentials, respectively. Using the automatic sorting procedure described above, we find that the correlation coefficients reach a value of 0.85 for a subset of 59 mutations for the C^{μ}–C^{μ} potential and of 62 mutations for the C^{β}–C^{β} potential. However, the excluded mutations do not seem to have common characteristics, and are moreover not the same for the C^{μ}–C^{μ} and C^{β}–C^{β} potentials, suggesting that these high correlations may not be physically relevant.

The residue–residue potentials appear thus to be less well suited than the backbone torsion potentials for estimating the stability change upon mutation of solvent-accessible residues. The reason for this can be understood by considering the type of interactions that both potentials describe. Indeed, though the residue–residue potentials have a component that represents local interactions along the chain (see Methods), they are completely dominated by hydrophobic interactions (Casari & Sippl, 1992; Bryant & Lawrence, 1993; Kocher *et al*., 1994), which are clearly not the dominant interactions at the protein surface. At the surface, the local interactions along the chain described by the backbone torsion potentials appear to have a more important role. To confirm these conclusions further, we compare the $\Delta\Delta G$ values with the transfer energies.

Relation Between $\Delta\Delta G$ Values and Transfer Energies

It has been pointed out that the transfer energies of residues from water to organic solvent (Nozaki & Tanford, 1971) correlate well with experimentally measured $\Delta\Delta G$ values obtained upon substitution of an amino acid residue buried in the protein core for all other amino acids (Yutani *et al*., 1987). This result stresses the importance of the hydrophobic effect in the protein core. At the protein surface, however, the measured $\Delta\Delta G$ values do not correlate with transfer energies, and hydrophobicity has thus only a marginal effect on protein stability. We find indeed that, for the mutation of Ala32 in barnase,

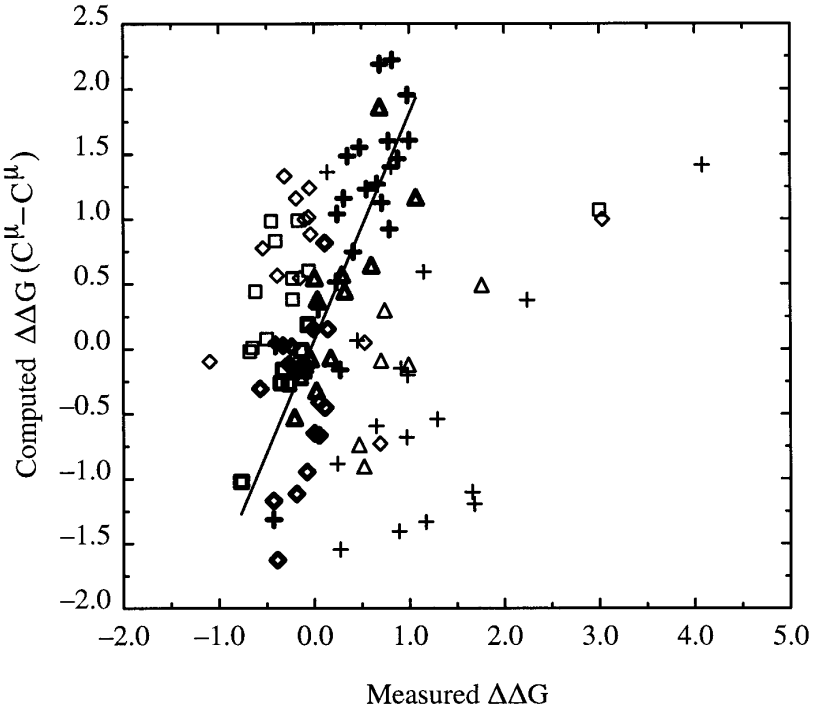

Figure 4. ΔΔ*G* values computed with the C^{μ}–C^{μ} potential as a function of the measured ΔΔ*G* values, for the 106 mutants listed in Table 1A. The measured ΔΔ*G* values are taken from O'Neil & DeGrado (1990), Serrano *et al*. (1990, 1992a,b), Matoushek *et al*. (1989), Sali *et al*. (1991), Horovitz *et al*. (1992), Dao-Pin *et al*. (1990), Hu *et al*. (1992), Blaber *et al*. (1993), Jackson & Fersht (1994), Itzhaki *et al*. (1995) and Zhang *et al*. (1995). The correlation coefficient on the 106 mutants is equal to 0.21 ($\mathscr{P}$ = 0.028241). When applying our sorting procedure, the following 47 mutations are rejected: Thr16 → Ser, Thr16 → Gly, Asp8 → Gly, Asp8 → Ala, Asp12 → Gly, Asp8 → Ser, Thr105 → Val, Thr16 → Ala, Ala32 → Pro, Gln31 → Gly, Asp12 → Ser, Ala32 → Gly, Asp12 → Ala, Ile55 → Ala, Ser28 → Gly and Ala32 → Arg in barnase; Ser44 → Ile, Ser44 → Pro, Ser44 → Gly, Ser44 → Tyr, Ser44 → Trp, Ser44 → Leu, Ser44 → Val, Ser44 → Phe, Ser44 → His, Ala93 → Pro, Val131 → Gly, Val131 → Ile and Gly113 → Ala in T4 lysozyme; Gly14 → Pro, Gly14 → Trp, Gly14 → Leu, Gly14 → Arg, Gly14 → Met, Gly14 → Ile, Gly14 → His, Gly14 → Tyr, Gly14 → Lys, Gly14 → Cys, Gly14 → Val and Gly14 → Phe in the ODG peptide; Glu33 → Asp, Glu33 → Asn, Glu34 → Gln, Glu34 → Asp, Lys37 → Gly, Pro44 → Ala in chymotrypsin inhibitor 2. The 59 remaining mutations are indicated with symbols in bold. The correlation coefficient on these 59 mutations is equal to 0.85 ($\mathscr{P}$ = 0.000000). The corresponding regression line is depicted. See the legend to Figure 1 for further details.

Ser44 in T4 lysozyme and Gly14 in the ODG peptide to the ten amino acids for which transfer energies are available (Nozaki & Tanford, 1971), the correlation coefficients between measured ΔΔ*G* values and transfer energies are 0.23, –0.28 and –0.12, respectively.

On the other hand, the ΔΔ*G* values computed with C^{μ}–C^{μ} potentials correlate rather well with the transfer energies (Figure 5), with correlation coefficients of 0.81, 0.67 and 0.75 for the mutation of Ala32 in barnase, Ser44 in T4 lysozyme and Gly14 in the ODG peptide, respectively. This is in agreement with the fact that these potentials are dominated by hydrophobic interactions (Casari & Sippl, 1992; Bryant & Lawrence, 1993; Kocher *et al*., 1994). The C^{β}–C^{β} potentials yield lower coefficients of 0.67, 0.49 and 0.26. On the contrary, the ΔΔ*G* values computed with the four types of backbone torsion potentials do not correlate at all with transfer energies. They even display an anti-correlation, with negative correlation coefficients between –0.1 and –0.5, according to the torsion potential. Thus, these potentials do not describe hydrophobic interactions at all, but rather local interactions along the sequence.

Conclusion

The main result of this study is that backbone torsion potentials can be used to predict reliably the stability change upon mutation of residues that are solvent-accessible, as monitored by a correlation coefficient of 0.86 for a total of 135 mutations. It must be stressed that these mutations are introduced at 45 different sites on six different proteins and peptides. The computed ΔΔ*G* values can thus be compared among mutations at different sites and in different proteins. The only limitation of our approach is that it is restricted to single-site mutations that do not affect the backbone structure of the wild-type; more precisely, for which the wild-type and mutant structures have their residues in the same torsion angle domains. This success is especially striking as the potentials used are derived from known protein structures and involve a lot of approximations (Rooman & Wodak, 1995). This results from the use of potentials derived from backbone torsion angle propensities, which appear to describe quite reliably the dominant interactions at the protein surface. These potentials have also been found to be adequate for identifying peptides with preferred conformation in solution as well as protein segments with well-defined conformations in absence of tertiary interactions, which could correspond to early folding intermediates (Rooman *et al*., 1992; Rooman & Wodak, 1992). Note that the folding free energies derived from these potentials cannot be directly compared with the energies obtained from semi-empirical force fields. They correspond to a particular combination of entropic

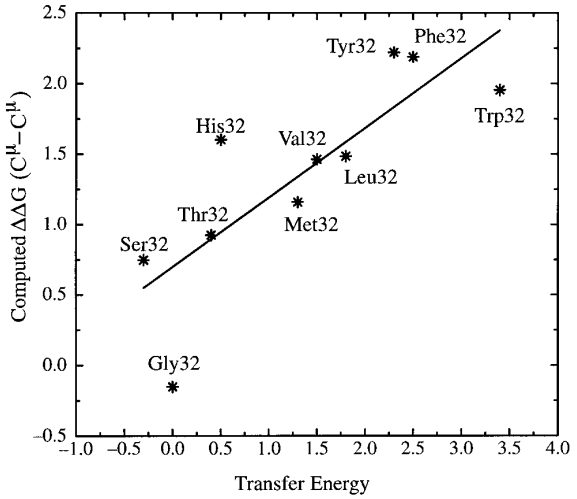

Figure 5. $\Delta\Delta G$ values computed with the C^{μ}–C^{μ} potential for the mutations at position 32 in barnase as a function of the transfer free energy from water to organic solvent (Nozaki & Tanford, 1971). The transfer energies are available for 11 of the amino acids. As one of them corresponds to the mutated residue (Ala), the correlation is only for ten amino acids. The names of these ten amino acids are indicated. The linear correlation coefficient is equal to 0.81 ($\mathscr{P}$ = 0.004441).

contributions and several energy terms, including van der Waals and electrostatic terms, which can be seen as being at the basis of the formation of local (secondary) structure.

Furthermore, the results of the present analysis justify *a posteriori* the formalism that is used for deriving potentials from structure data, given in equation (2). Indeed, our whole analysis rests on the assumptions that these potentials yield folding free energies and that the considered denatured-like state is a good approximation of the true. Only when the denatured states of the wild-type and the mutant are significantly different, as it is the case for one of the chemotactic protein mutations, does our approximation seem to break down. However, though the present results clearly support the correctness of the formalism used, they do not rigorously prove it, as we have no idea of the sensitivity of the results on the chosen formalism. To examine this issue further, additional tests must be performed. For that purpose, other definitions of denatured-like states, in particular those proposed by Rooman & Wodak (1995), will be tested for evaluating changes in folding free energies and the resulting effect on the correlation with measured stability changes will be examined. The analysis will also be extended to the mutations of residues that are buried in the protein core.

Methods

Backbone torsion potentials

Backbone torsion potentials consider only interactions between neighboring residues along the chain (Rooman *et al*., 1991, 1992; Kocher *et al*., 1994). They are based on the propensities of residues to be associated with certain values of the backbone torsion angles (φ, ψ, ω). For that purpose, the (φ, ψ, ω) map (Ramachandran & Sasisekharan, 1968) is divided into seven torsion angle domains, six for the *trans* conformation, denoted A, C, B, P, E and G, and one for the *cis* conformation, denoted O (Rooman *et al*., 1991, 1992). A corresponds to right-handed α-helical structures, C to 3_{10}-helices, B and P to extended structures, with B being characteristic of β-strands, G to left-handed helices and E to left-handed extended conformations.

Two types of backbone torsion potentials are considered. The residue → torsion potential (Rooman *et al*., 1991, 1992; Kocher *et al*., 1994) takes into account the probability that a residue a_i at position i along the sequence, and pairs of residues (a_i, a_j) at positions i and j along the sequence, are associated with a torsion angle domain t_k at position k. Equation (2) becomes in this case:

$$\Delta G^S(C) = -kT \sum_{i,j,k=1}^{N} \frac{1}{\zeta_k} \log \frac{P(t_k|a_i, a_j)}{P(t_k)} \qquad (3)$$

where N is the number of residues in the sequence S. We consider a short range backbone potential, which comprises contributions from residues in the interval $k-1 \leqslant i \leqslant j \leqslant k+1$ along the sequence, and a middle range potential, with $k-8 \leqslant i \leqslant j \leqslant k+8$. The normalization factor ζ_k ensures that the contribution of each residue in the window $[k-1, k+1]$ or $[k-8, k+8]$ is counted once. It is equal to the window width, except near chain ends.

The torsion → residue potential (Kocher *et al*., 1994) is in some sense the converse of the above potential, and takes into account correlations between torsion angle domains rather than between amino acids. It considers the probability that the torsion angle domain t_i at position i along the sequence, and pairs of domains (t_i, t_j) at positions i and j along the sequence, are associated with an amino acid a_k at position k. Equation (2) becomes in this case:

$$\Delta G^S(C) = -kT \sum_{i,j,k=1}^{N} \frac{1}{\zeta_k} \log \frac{P(t_i, t_j|a_k)}{P(t_i, t_j)} \qquad (4)$$

where N is the number of residues in the sequence S. The short range and middle range potentials are defined as above.

Residue–residue potentials

Residue–residue interaction potentials describe both local interactions along the chain and interactions between residues that are far apart along the sequence but close in space. They are computed from propensities of two residues a_i and a_j, at positions i and j along the sequence, to be separated by a spatial distance d_{ij} (Kocher *et al*., 1994). Consecutive residues along the chain are not considered. Probabilities of residues separated by one to seven positions along the sequence are computed separately, whereas probabilities of residues separated by eight positions and more are all merged. This distinction yields a potential that represents both local and non-local interactions along the chain. The folding energy defined by these potentials is, according to equation (2):

$$\Delta G^S(C) = -kT \sum_{i,j=1}^{N} \log \frac{P^{|i-j|}(d_{ij}|a_i, a_j)}{P^{|i-j|}(d_{ij})} \qquad (5)$$

with $i+1 < j$ and with the probabilities $P^{|i-j|}$ being independent of $|i-j|$ for $|i-j| > 8$. The inter-residue

distances d_{ij} are computed either between the C^{β}-atoms or between the average centroïds C^{μ}. These centroïds are specific for each amino acid type and are defined as the average of the atomic co-ordinate centers of all conformations of side-chains of the same type observed in the protein dataset (Kocher *et al*., 1994). These two types of residue–residue potentials are referred to as C^{β}–C^{β} and C^{μ}–C^{μ} potentials, respectively.

Correction for sparse data

When sequence elements are not represented often enough in the protein dataset, the computed frequencies may not be accurate. To correct for this, the conditional probabilities $P(c_i|s_j)$ in Equations (2) to (5) are replaced by linear combinations of $P(c_i|s_j)$ and of the sequence non-specific probabilities $P(c_i)$ (Sippl, 1990; Rooman *et al*., 1991, 1992; Kocher *et al*., 1994):

$$P(c_i|s_j) \rightarrow \frac{1}{\sigma + m_j^s}\left[\sigma P(c_i) + m_j^s P(c_i|s_j)\right] \qquad (6)$$

where m_i^s is the number of occurrences of the sequence element s_i and σ is a parameter. This expression ensures that the sequence non-specific probability dominates for unusual sequence patterns, and tends towards zero for frequent ones. Two values of σ were tested: $\sigma = 6$ and $\sigma = 50$. Though the ΔG values differ significantly according to which of the values is chosen, the correlations between the computed and measured differences in ΔG upon mutation are exactly the same. All the results presented in this paper are obtained with $\sigma = 50$.

Protein structure data

The mean force potentials are derived from a set of 141 well resolved ($\leqslant$2.5 Å) and refined proteins, with less than 20% sequence identity or no structural homology (Lemer, C., Rooman, M. J. & Wodak, S. J., unpublished results). Information on the proteins is extracted from the protein database SESAM (Huysmans *et al*., 1991), which contains sequence and structure information on proteins from the Brookhaven databank (Bernstein *et al*., 1977).

Computing folding free energies for wild-type and mutant proteins

To estimate the stability of wild-type *versus* mutant proteins, their respective folding free energies, $\Delta G^{wild\text{-}type}(C)$ and $\Delta G^{mutant}(C)$, are computed using equations (2) to (5), where C corresponds to the native structure of the wild-type. It is thus assumed that the wild-type and mutant proteins have the same backbone structure. The difference in folding free energy, $\Delta\Delta G$, between mutant and wild-type, is estimated using the following sign convention:

$$\Delta\Delta G = \Delta G^{mutant}(C) - \Delta G^{wild\text{-}type}(C) \qquad (7)$$

The folding free energy difference is thus negative when the mutant protein is more stable than the wild-type protein.

To avoid biasing the potentials towards the native structure, the jack-knife procedure is applied. It consists of excluding from the dataset used to compile the statistics the proteins that display sequence homology with the protein under study. The homology criterion used for this purpose is the same as the one used for defining the protein dataset (see above). It turned out that the $\Delta\Delta G$ values computed with the potentials considered in this paper are insensitive to the jack-knife procedure, up to the second decimal, when a value of $\sigma = 50$ is used. This is due to the fact that, for all considered mutations, the residue pairs comprising the mutated residue are sufficiently represented in our dataset to yield reliable statistics.

Correlating measured and computed $\Delta\Delta G$ values

The computed $\Delta\Delta G$ values are compared with the experimentally determined values, and the correlation coefficient is computed, assuming a linear regression. To estimate the significance of this correlation, the probability, referred to as $\mathscr{P}$, that the same correlation would arise by random sampling in an uncorrelated population is computed (Fisher, 1958).

Acknowledgements

We are grateful to J.-P. Kocher, M. Prévost, D. Van Belle & S. Wodak for helpful discussions. D.G. is a research assistant at the Fonds pour la Formation à la Recherche dans l'Industrie et l'Agriculture (FRIA). M.R. is a research associate at the Belgian National Fund for Scientific Research (FNRS). We also thank the Belgian Program of Interuniversity Poles of Attraction (PAI) for support.

References

Alard, P. (1991). Calculs de surface et d'énergie dans le domaine des macromolécules, PhD thesis, Université Libre de Bruxelles.

Basch, P. A., Singh, U. C., Langridge, R. & Kollman, P. A. (1987). Free energy calculations by computer simulation. *Science*, **236**, 564–568.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meywe, E. F., Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanoushi, T. & Tasumi, M. (1977). The protein data bank: a computer-based aerchival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–542.

Blaber, M., Zang, X. & Matthews, B. W. (1993). Structural basis of amino acid α helix propensity. *Science*, **260**, 1637–1640.

Bryant, S. H. & Lawrence, C. E. (1993). An empirical energy function for threading protein sequence through folding motif. *Proteins: Struct. Funct. Genet.* **16**, 92–112.

Casari, G. & Sippl, M. J. (1992). Structure-derived hydrophobic potential. Hydrophobic potential derived from X-ray structures of globular proteins is able to identify native folds. *J. Mol. Biol.* **224**, 725–732.

Dang, L. X., Merz, K. M., Jr & Kollman, P. A. (1989). Free energy calculations on protein stability: Thr157 → Val157 mutation of T4 lysozyme. *J. Am. Chem. Soc.* **111**, 8505–8508.

Dao-Pin, S., Baase, W. A. & Matthews, B. W. (1990). A mutant T4 lysozyme (Val131 → Ala) designed to increase thermostability by the reduction of strain within an α-helix. *Proteins: Struct. Funct. Genet.* **7**, 198–204.

Eriksson, A. E., Baase, W. A., Zhang X.-J., Heinz, D. W., Blaber, M., Baldwin, E. P. & Matthews, B. W. (1992). Response of a protein structure to cavity-creating mutations and its relation to the hydrophobic effect. *Science*, **255**, 178–183.

Fersht, A. R. & Serrano, L. (1993). Principles of protein

stability derived from protein engineering experiments. *Curr. Opin. Struct. Biol.* **3**, 75–83.

Fisher, R. A. (1958). *Statistical Methods for Research Workers*, Oliver and Boyd Ltd, Edinburgh.

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.

Horovitz, A., Matthews, J. M. & Fersht, A. R. (1992). α-Helix stability in proteins. II. Factors that influence stability at an internal position. *J. Mol. Biol.* **227**, 560–568.

Hu, C.-Q., Kitamura, S., Tanaka, A. & Sturtevant, J. M. (1992). Differential scanning calorimetric study of the thermal unfolding of mutant forms of phage T4 lysozyme. *Biochemistry*, **31**, 1643–1647.

Huysmans, M., Richelle, J. & Wodak, S. J. (1991). SESAM, a relational database for structure and sequence of macromolecules. *Proteins: Struct. Funct. Genet.* **11**, 59–76.

Hynes, T. R., Hodel, A. & Fox, R. O. (1994). Engineering alternative β-turn types in staphylococcal nuclease. *Biochemistry*, **3**, 5021–5030.

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.

Jackson, S. E. & Fersht, A. R. (1994). Contribution of residues in the reactive site loop of chymotrypin inhibitor 2 to protein stability and activity. *Biochemistry*, **33**, 13880–13887.

Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Kellis, J. T., Jr, Nyberg, K. & Fersht, A. R. (1989). Energetics of complementary side-chain packing in a protein hydrophobic core. *Biochemistry*, **28**, 4914–4922.

Kocher, J.-P. A., Rooman, M. J. & Wodak, S. J. (1994). Factors influencing the ability of knowledge-based potentials to identify native sequence-structure matches. *J. Mol. Biol.* **235**, 1598–1613.

Koehl, P. & Delarue, M. (1994). Polar and nonpolar atomic environments in the protein core: implications for folding and binding. *Proteins: Struct. Funct. Genet.* **20**, 264–278.

Lee, C. (1994). Predicting protein mutant energetics by self-consistent ensemble optimization. *J. Mol. Biol.* **236**, 918-939.

Lee, C. & Levitt, M. (1991). Accurate prediction of the stability and activity effects on site-directed mutagenesis on a protein core. *Nature*, **352**, 448–451.

López-Hernández, 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.

Matoushek, A., Kellis, J. T., Jr, Serrano, L. & Fersht, A. R. (1989). Mapping the transition state and pathway of protein folding by protein engineering. *Nature*, **340**, 122–126.

Minor, D., Jr & Kim, P. S. (1994). Context is a major determinant of β-sheet propensity. *Nature*, **371**, 264–267.

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

Miyazawa, S. & Jernigan, R. L. (1994). Protein stability for single substitution mutants and the extent of local compactness in the denatured state. *Protein Eng.* **7**, 1209–1220.

Muñoz, V. & Serrano, L. (1994). Intrinsic secondary structure propensities of the amino acids, using statistical f-y matrices: comparison with experimental data. *Proteins: Struct. Funct. Genet.* **20**, 301–311.

Muñoz, V., López, 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*, **3**, 5858–5866.

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.

O'Neil, K. T. & DeGrado, W. F. (1990). A thermodynamic scale for the helix-forming tendencies of the commonly occurring amino acids. *Science*, **250**, 646–650.

Piela, L., Nemethy, G. N. & Scheraga, H. A. (1987). Proline-induced constraints in α-helices. *Biopolymers*, **26**, 1587–1600.

Pohl, F. M. (1971). Empirical protein energy maps. *Nature New Biol.* **237**, 277–279.

Ramachandran, G. & Sasisekharan, V. (1968). Conformation of polypeptides and proteins. *Advan. Protein Chem.* **23**, 283–437.

Richardson, J. S. & Richardson, D. C. (1988). Amino acid preferences for specific locations at the ends of a helices. *Science*, **240**, 1648–1652.

Rooman, M. J. & Wodak, S. J. (1992). Extracting information on folding from the amino acid sequence: role of consensus stable regions in homologous proteins. *Biochemistry*, **31**, 10239–10249.

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.

Rooman, M. J., Kocher, J.-P. A. & Wodak, S. J. (1991). Prediction of protein backbone conformation based on 7 structure assignments: influence of local interactions. *J. Mol. Biol.* **221**, 961–979.

Rooman, M. J., Kocher, J.-P. A. & Wodak, S. J. (1992). Extracting information on folding from the amino acid sequence: accurate predictions for protein regions with stable conformation in absence of tertiary interactions. *Biochemistry*, **31**, 10226–10238.

Rose, G. D., Geselowitz, A. R., Lesser, G. J., Lee, R. H. & Zehfus, M. H. (1985). Hydrophobicity of amino acid residues in globular proteins. *Science*, **29**, 834–838.

Sali, D., Bycroft, M. & Fersht, A. R. (1991). Surface electrostatic interactions contribute little to stability of barnase. *J. Mol. Biol.* **220**, 779–788.

Serrano, L., Horovitz, A., Avron, B., Bycroft, M. & Fersht A. R. (1990). Estimating the contribution of engineered surface electrostatic interactions to protein stability by using double-mutant cycles. *Biochemistry*, **29**, 9343–9352.

Serrano, L., Sancho J., Hirshberg, M. & Fersht, A. R. (1992a). α-Helix stability in proteins. I. Empirical correlations concerning substitution of side-chains at the N and C-caps and the replacement of alanine by glycine or serine at solvent-exposed surfaces. *J. Mol. Biol.* **227**, 544–559.

Serrano, L., Kellis, J. T., Jr, Cann, P., Matouschek, A. & Fersht, A. R. (1992b). The folding of an enzyme. II.

Substructure of barnase and the contribution of different interactions to protein stability. *J. Mol. Biol.* **224**, 783–804.

Shoichet, B. K., Baase, W. A., Kuroki, R. & Matthews, B. W. (1995). A relationship between protein stability and protein function. *Proc. Natl Acad. Sci. USA*, **92**, 452–456.

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.

Sippl, M. J. (1990). Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge based prediction of local structures in globular proteins. *J. Mol. Biol.* **213**, 859–883.

Sippl, M. J. (1993). Boltzmann's principle, knowledge-based mean fields and protein folding, An approach to the computational determination of protein structures. *J. Comp. Aided Mol. Design* **7**, 473–501.

Sippl, M. J. (1995). Knowledge-based potentials for proteins. *Curr. Opin. Struct. Biol.* **5**, 229–235.

Tidor, B. & Karplus, M. (1991). Simulation analysis of the stability mutant R96H of T4 lysozyme. *Biochemistry*, **30**, 3217–3228.

van Gunsteren, W. F. & Mark, A. E. (1992). Prediction of the activity and stability effects of site-directed mutagenesis on a protein core. *J. Mol. Biol.* **227**, 389–395.

Wintjens, R. T., Rooman, M. J. & Wodak, S. J. (1996). Automatic classification and analysis of αα-turn motifs in proteins. *J. Mol. Biol.* **255**, 235–253.

Wodak, S. J. & Rooman, M. J. (1993). Generating and testing protein folds. *Curr. Opin. Struct. Biol.* **3**, 247–259.

Yutani, K., Ogasahara, K., Tsujita, T. & Sugino, Y. (1987). Dependence of conformational stability on hydrophobicity of the amino acid residue in a series of variant proteins substituted at a unique position of tryptophan synthase α-unit. *Proc. Natl Acad. Sci. USA*, **84**, 4441–4444.

Zhang, X., Baase, W. A., Shoichet, B. K., Wilson, K. P. & Matthews, B. W. (1995). Enhancement of protein stability by the combination of point mutation in T4 lysozyme is additive. *Protein Eng.* **8**, 1017–1022.

Edited by R. Huber

(Received 3 November 1995; received in revised form 15 January 1996; accepted 25 January 1996)

Note added in proof: We verified that this is the case on the structure of the Ala82 → Pro mutant (1L24).