J. Mol. Biol. (1997) **272**, 276–290

JMB

Predicting Protein Stability Changes upon Mutation Using Database-derived Potentials: Solvent Accessibility Determines the Importance of Local *versus* Non-local Interactions Along the Sequence

Dimitri Gilis* and Marianne Rooman

UCMB, Université Libre de Bruxelles, CP160/16 av. F. Roosevelt 50 1050 Brussels, Belgium

**Corresponding author*

For 238 mutations of residues totally or partially buried in the protein core, we estimate the folding free energy changes upon mutation using database-derived potentials and correlate them with the experimentally measured ones. Several potentials are tested, representing different kinds of interactions. Local interactions along the chain are described by torsion potentials, based on propensities of amino acids to be associated with backbone torsion angle domains. Non-local interactions along the sequence are represented by distance potentials, derived from propensities of amino acid pairs or triplets to be at a given spatial distance. We find that for the set of totally buried residues, the best performing potential is a combination of a distance potential and a torsion potential weighted by a factor of 0.4; it yields a correlation coefficient between computed and measured changes in folding free energy of 0.80. For mutations of partially buried residues, the best potential is a combination of a torsion potential and a distance potential weighted by a factor of 0.7, and for the previously analysed mutations of solvent accessible residues, it is a torsion potential taken individually; the respective correlation coefficients reach 0.82 and 0.87. These results show that distance potentials, dominated by hydrophobic interactions, represent best the main interactions stabilizing the protein core, whereas torsion potentials, describing local interactions along the chain, represent best the interactions at the protein surface. The prediction accuracy reached by the distance potentials is, however, lower than that of the torsion potentials. A possible reason for this is that distance potentials would not describe correctly the effect on protein stability due to cavity formation upon mutating a large into a small amino acid. Last but not least, our results indicate that although local interactions, responsible for secondary structure formation, do not dominate in the protein core, they are not negligible for all that. They have a significant weight in the delicate balance between all the interactions that ensure protein stability.

Keywords: single-site mutations; folding free energies; protein stability; mean force potentials

Introduction

For more than 20 years, experimenters and theorists have tried to understand what kind of interactions govern the first stages of protein folding and lead to the formation of folding intermediates, and what are the forces that maintain protein stability (Gō & Taketomi, 1978; Govindarajan & Goldstein, 1995; Unger & Moult, 1996; Shakhnovich *et al*., 1996; Fersht, 1995; Muñoz & Serrano, 1996). The main questions concern the relative importance of hydrophobic *versus* more specific interactions, and of local *versus* non-local interactions along the sequence. Site-directed mutagenesis appears to be a powerful tool to study interactions both in native proteins and in folding intermediates (Matouschek *et al*., 1989; Serrano *et al*., 1992a; Fersht *et al*., 1992; Itzhaki *et al*., 1995). It is indeed possible to detect stability changes caused by mutation, in the folded and transition

0022–2836/97/370276–15 $25.00/0/mb971237

states, by measuring and comparing the changes in unfolding and activation free energies. Doing so, one can get information about the structures that are already formed in the transition state and, as a consequence, about the interactions that drive the folding process.

Whereas now there exists a lot of experimental data on folding free energy changes upon mutation obtained by site-directed mutagenesis experiments (see references in legend to Table 1), only a few theoretical methods have been developed to predict such stability changes. Some of these methods are based on detailed atomic models coupled to semi-empirical force fields (Basch *et al.*, 1987; Tidor & Karplus, 1991) and others on rougher descriptions of protein structure or on simplified energetic criteria (Lee & Levitt, 1991; Lee, 1994; Koehl & Delarue, 1994; Muñoz & Serrano, 1994; Miyazawa & Jernigan, 1994; Sippl, 1995). Their performances are, in general, evaluated by comparing the calculated folding free energies to the measured ones and are reasonably good. However, it must be stressed that the performance tests are restricted to selected mutations in a single protein, usually even at a single site. In most studies, the mutated residues are buried in the protein core; since hydrophobic interactions dominate in these regions, the energetic criteria obviously involve hydrophobicity. In the few studies analysing mutations of solvent accessible residues, the stability changes are correlated with statistical propensities of single amino acids to be in α-helices or β-strands (Muñoz & Serrano, 1994), or with distance-dependent residue-residue potentials (Sippl, 1995).

A previous paper of ours (Gilis & Rooman, 1996) was also devoted to the prediction of stability changes upon mutation of solvent accessible residues, with the notable difference that we merged mutations at different sites and in different proteins. We showed that database-derived potentials based on backbone torsion angle propensities can predict reliably the stability changes upon mutation of residues that have a solvent accessibility larger than 50%. These potentials describe local interactions along the sequence; they do not take into account the spatial environment of the residues, but they do consider their environment along the sequence. For 96 out of the 106 considered mutations, introduced in four different proteins and peptides, a quite good correlation between computed and measured changes in folding free energies has been obtained, as measured by a correlation coefficient of 0.87. The ten excluded mutations seemed to involve modifications of the native or denatured backbone structures (we were able to verify this for one of the mutants whose structure has been determined), thereby contradicting the basic assumptions of our approach, or to involve strong hydrophobic interactions, which are atypical for surface residues. These results led to the conclusion that local interactions along the sequence dominate at the surface of the protein.

This analysis is extended and completed here, with the aim of determining the relative weight of different types of interactions in each protein region, and in particular of local *versus* non-local interactions along the chain. For that purpose, single-site mutations involving residues with a solvent accessibility of less than 50% are collected. The changes in folding free energy are estimated using several database-derived potentials, describing different types of interactions. These potentials are combined with relative weighting coefficients. By determining the values of these coefficients that yield the best correlations between experimentally measured and computed changes in folding free energy, information is obtained about the interactions that ensure protein stability.

Ensemble of Mutations Considered

A total of 238 mutants is considered (Table 1), whose stability relative to the wild-type has been measured experimentally. They correspond to mutations of residues buried in the protein core, exposing at most 50% of their accessible surface to the solvent as calculated by SurVol (Alard, 1991) They involve only single amino acid substitutions that are not supposed (by their authors) to perturb the structure of the native and denatured states. The reason for the latter restriction is that the mutants are given the same backbone atomic coordinates as the wild-type (see equation (5)). So, we do not take into account structural rearrangements upon mutation.

The 238 mutations are introduced at 107 different sites in seven proteins: human, chicken, and T4 lysozyme, barnase, tryptophan synthase, chymotrypsin inhibitor 2 and apomyoglobin. They are introduced in all types of secondary structures. In 121 of them, the mutated residue has a solvent accessibility comprised between 0 and 20% (Table 1A), in 69 the accessibility is between 20 and 40% (Table 1B), and in 48 it is between 40 and 50% (Table 1C).

Mutations of Residues with a Solvent Accessibility of Less than 20%

The change in folding free energy $\Delta\Delta G$ between wild-type and mutant proteins is estimated using database-derived potentials, either alone or in combination, for the 121 mutants of completely buried residues listed in Table 1A. The different kinds of potentials are described in Methods. They involve on the one hand distance potentials, in particular the C^{μ}–C^{μ} potential, based on propensities of amino acid pairs to be separated by a certain spatial distance measured between average side-chain centroids C^{μ}, and the C^{μ}–$C^{\mu}_{\text{long-range}}$ potential, where only residue pairs separated by more than 15 residues along the sequence are taken into account in the statistics. On the other hand, the torsion potentials $\text{torsion}_{\text{short-range}}$ and

Table 1. Set of 238 considered mutations

Protein[a]	Sequence position	Mutated residues	Mutant residues	(ϕ,ψ,ω) Domain[b]	Secondary structure[c]	Solvent accessibility[d]
A. *Set of 121 mutations of totally buried residues, with solvent accessibility of less than 20%*[e]						
Barnase	5	Asn	Ala	C	C	13.8
(1RNB)	7	Phe	Leu	A	H	10.2
	10	Val	Thr,Ala	A	H	0.0
	13	Tyr	Ala	A	H	20.0
	14	Leu	Ala	A	H	0.0
	24	Tyr	Phe	B	E	9.0
	25	Ile	Val,Ala	B	E	8.0
	41	Asn	Asp	C	C	14.5
	51	Ile	Ala,Val	P	E	1.1
	54	Asp	Ala,Asn	P	E	20.0
	58	Asn	Ala,Asp	G	C	4.5
	76	Ile	Val,Ala	B	C	0.0
	78	Tyr	Phe	B	C	3.6
	88	Ile	Val,Ala	B	E	0.0
	89	Leu	Val	B	E	0.0
	91	Ser	Ala	B	E	1.3
	96	Ile	Val,Ala	B	E	0.0
	99	Thr	Val	B	E	0.5
	103	Tyr	Phe	G	S	13.9
T4 lysozyme (1LYD)	3	Ile	Trp,Tyr,Phe, Leu,Val,Met, Cys,Ala,Thr Ser,Gly,Glu, Asp	A	H	12.2
	30	Gly	Ala,Phe	G	T	19.8
	77	Gly	Ala	A	H	14.9
	98	Ala	Val	A	H	0.0
	117	Ser	Val,Ile,Phe	A	H	0.0
	133	Leu	Ala	A	H	0.5
	149	Val	Cys	A	H	0.0
	152	Thr	Ser	A	H	0.0
Human lysozyme (1LZ1)	23	Ile	Val	B	B	7.0
	56	Ile	Val	C	T	0.4
	59	Ile	Val	P	E	0.9
	89	Ile	Val	C	C	2.8
	106	Ile	Val	C	G	2.4
Trp synthase (1WSY)	49	Glu	Gly,Ala,Val, Ile,Leu,Pro, Tyr,Phe,Trp, His,Lys,Asn, Gln,Asp,Cys, Met,Thr,Ser	B	E	1.1
Chymotrypsin inhibitor 2 (2CI2)	27	Leu	Ala	C	G	0.6
	31	Ser	Gly,Ala	P	B	16.1
	35	Ala	Gly	A	H	0.0
	39	Ile	Val,Leu	A	H	0.0
	43	Lys	Ala,Gly	B	C	2.7
	51	Leu	Ala,Ile,Val	B	E	19.6
	66	Val	Ala	B	E	0.0
	67	Arg	Ala	P	E	15.9
	68	Leu	Ala	B	E	0.0
	69	Phe	Leu,Val,Ala	B	E	15.2
	70	Val	Ala	B	E	7.6
	76	Ile	Val,Ala	P	B	1.5
	80	Pro	Ala	P	C	0.9
	82	Val	Thr,Ala,Gly	P	E	12.9
Chicken lysozyme (4LYZ)	3	Phe	Tyr	P	C	7.7
	15	His	Leu	C	T	20.0
	31	Ala	Val,Ile,Leu	A	H	0.0
	40	Thr	Ser,Ile	C	S	0.0
	55	Ile	Leu,Val,Phe Ala,Thr	A	T	1.0
	91	Ser	Thr,Val,Ala, Asp,Tyr	A	H	0.1

Table 1—*Continued*

Protein[a]	Sequence position	Mutated residues	Mutant residues	(ϕ,ψ,ω) Domain[b]	Secondary structure[c]	Solvent accessibility[d]
Apomyoglobin	7	Trp	Phe	A	H	4.8
(4MBN)	14	Trp	Phe	A	H	0.7
	68	Val	Thr	A	H	1.5
	123	Phe	Lys	B	S	1.1
	130	Ala	Lys,Leu	A	H	1.7
	131	Met	Ala	A	H	0.0
B. *Set of 69 mutations of residues with a solvent accessibility between 20 and 40%*[f]						
Barnase	4	Ile	Val,Ala	B	C	28.1
(1RNB)	26	Thr	Ser,Val,Ala, Gly,Asn,Gln, Glu,Asp	P	C	24.2
	27	Lys	Gly	A	H	29.5
	34	Gly	Ala,Ser,Asn, Asp,His,Lys, Arg,Thr	G	T	36.1
	45	Val	Ala,Thr	C	H	28.1
	62	Lys	Arg	C	S	31.8
	84	Asn	Ala	B	C	25.1
	109	Ile	Val,Ala	A	C	32.1
	110	Arg	Ala	A	C	30.5
T4 lysozyme	11	Glu	Phe,Met,Ala	A	H	22.5
(1LYD)	38	Ser	Asp	B	C	36.2
	41	Ala	Val	A	H	27.9
	132	Asn	Met,Phe,Ile	A	H	28.2
	157	Thr	Val,Asn,Ser, Asp,Gly,Cys, Leu,Arg,Ala, Glu,His,Phe, Ile	B	S	30.8
Chymotrypsin	21	Lys	Ala,Met	B	C	34.7
inhibitor 2	22	Thr	Val,Ala,Gly	C	C	35.6
(2CI2)	30	Lys	Ala	P	S	34.5
	38	Val	Ala	A	H	31.6
	48	Ile	Val,Ala	P	E	21.2
	55	Thr	Ser,Val,Ala	P	C	37.4
	57	Val	Ala	B	C	36.9
	64	Asp	Ala	C	T	20.7
	71	Asp	Ala	B	C	38.9
	75	Asn	Asp,Ala	B	B	25.6
	79	Val	Thr,Ala,Gly	P	C	24.6
Chicken lysozyme (4LYZ)	34	Phe	Tyr	C	H	33.0
Apomyoglobin (4MBN)	36	His	Gln	B	C	23.2
C. *Set of 48 mutations of residues with a solvent accessibility between 40 and 50%*[g]						
Barnase	6	Thr	Ser,Ala,Gly Asn,Asp,Gln, Glu,Pro	B	S	48.6
(1RNB)	17	Tyr	Ala,Gly,Ser	C	H	49.0
	18	His	Gln,Gly,Ala, Ser,Asn,Asp, Lys,Arg	G	S	46.5
	29	Glu	Gly,Ala,Ser	A	H	48.4
	33	Leu	Gln	C	T	48.7
	36	Val	Ala,Thr	B	C	41.6
	77	Asn	Ala	G	S	41.5
	92	Ser	Ala	A	T	43.2
T4 lysozyme	116	Asn	Asp	A	H	48.1
(1LYD)	128	Glu	Ala	A	H	43.9
Chymotrypsin	25	Pro	Ala	A	G	47.9
inhibitor 2	26	Glu	Ala,Gln	C	G	45.9
(2CI2)	36	Lys	Ala,Gly	A	H	41.6

Table 1—*Continued*

Protein[a]	Sequence position	Mutated residues	Mutant residues	(ϕ,ψ,ω) Domain[b]	Secondary structure[c]	Solvent accessibility[d]
	40	Leu	Ala,Gly	C	H	41.5
	42	Asp	Ala	A	H	41.8
	49	Ile	Val,Ala,Gly Thr	B	E	40.0
	53	Val	Thr,Ala,Gly	P	T	48.6
	60	Glu	Ala	B	C	49.7
	62	Arg	Ala	B	C	44.4
	77	Ala	Gly	C	C	40.7
Chicken lysozyme (4LYZ)	68	Arg	Lys	C	S	44.8

[a] The codes in parentheses correspond to the PDB codes (Bernstein *et al*., 1977) of the proteins.
[b] The (ϕ,ψ,ω) domains of the mutated residues in the wild-type structure, using definitions of Rooman *et al*. (1991). 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.
[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).
[e] The measured ΔΔGs are taken from Serrano *et al*. (1992a), Matouschek *et al*. (1989), Kellis *et al*. (1988) Matsumura *et al*. (1988), Eriksson *et al*. (1992), Shoichet *et al*. (1995), Matthews *et al*. (1987, 1993), Daopin *et al*. (1991), Jackson *et al*. (1993), Itzhaki *et al*. (1995), Otzen & Fersht (1995), Takano *et al*. (1995), Yutani *et al*. (1987), Shih *et al*. (1995), Shih & Kirsch (1995) and Kay & Baldwin (1996).
[f] The measured ΔΔGs are taken from Serrano *et al*. (1992a,b), Serrano & Fersht (1989), Zhang *et al*. (1995), Daopin *et al*. (1990), Alber *et al*. (1987), Shoichet *et al*. (1995), Otzen & Fersht (1995), Itzhaki *et al*. (1995), Jackson *et al*. (1993), Shih & Kirsch (1995) and Kay & Baldwin (1996).
[g] The measured ΔΔGs are taken from Serrano *et al*. (1992a,b), Matouschek *et al*. (1989), Serrano & Fersht (1989), Jackson & Fersht (1994), Itzhaki *et al*. (1995), Otzen & Fersht (1995), Zhang *et al*. (1992, 1995) and Shih & Kirsch (1995).

$\text{torsion}_{\text{middle-range}}$ are based on propensities of amino acids to be associated with backbone torsion angle domains.

The correlation coefficient r between experimental ΔΔGs and ΔΔGs computed using the distance potential C^{μ}–C^{μ} is equal to 0.76, and is yet higher ($r = 0.78$) for the C^{μ}–$C^{\mu}_{\text{long-range}}$ potential, where purely non-local interactions along the sequence are taken into account. This increase is small but significant. Indeed, when removing 20 randomly chosen mutations from the total set of 121 mutations and computing the correlation coefficient on the 101 remaining mutations, higher values are obtained with the C^{μ}–$C^{\mu}_{\text{long-range}}$ than with the C^{μ}–C^{μ} potential in as much as 99.5% of the 1000 trials. The good performance of C^{μ}–C^{μ} potentials on mutations of buried residues is not surprising, since these potentials are dominated by hydrophobic interactions (Casari & Sippl, 1992; Bryant & Lawrence, 1993; Kocher *et al*., 1994), known to be the dominant forces in the protein core. The C^{μ}–$C^{\mu}_{\text{long-range}}$ potential, where the local interactions are completely cancelled, performs even better. In accordance with this result, the correlation coefficients obtained with the $\text{torsion}_{\text{short-range}}$ and $\text{torsion}_{\text{middle-range}}$ potentials, which describe purely local interactions along the sequence, are much lower (r equal to 0.42 and 0.46).

Additional information about the relative importance of the various interactions in the protein core is obtained by combining several potential terms with relative weighting coefficients (see Methods). The highest correlation coefficient on the set of 121 mutations is obtained with the sum of the C^{μ}–$C^{\mu}_{\text{long-range}}$ potential weighted by a factor of 1 and the $\text{torsion}_{\text{middle-range}}$ potentials weighted by a factor of 0.4; the correlation coefficient is then equal to 0.80 (Figure 1). Thus, the addition of a small contribution from the torsion potentials increases the correlation coefficient from 0.78 to 0.80. To verify the statistical significance of this small increase, we exclude again 20 randomly chosen mutations from the original set, and compute both the optimal weighting factors and the correlation coefficient on the remaining mutations; this procedure is repeated 1000 times. It is found that the correlation coefficients are always higher with the combined potentials than with the C^{μ}–$C^{\mu}_{\text{long-range}}$ potential alone, in each of the 1000 trials. These coefficients are equal, on the average, to those computed from the full set: 0.78 for the C^{μ}–$C^{\mu}_{\text{long-range}}$ potential taken individually and 0.80 for the optimal combination of the C^{μ}–$C^{\mu}_{\text{long-range}}$ and $\text{torsion}_{\text{middle-range}}$ potentials, with standard deviations of less than 0.02. Note that the optimal weighting factors are not the same in the 1000 trials, but on the average, they are equal to those obtained from the full set: keeping the weighting factor of the C^{μ}–$C^{\mu}_{\text{long-range}}$ potential equal to 1, the average factor of the $\text{torsion}_{\text{middle-range}}$ potential is equal to 0.4, with a standard deviation of 0.1.

To remove every residual doubt about the above results, an additional test is performed, showing that the 0.02 increase in the correlation coefficient

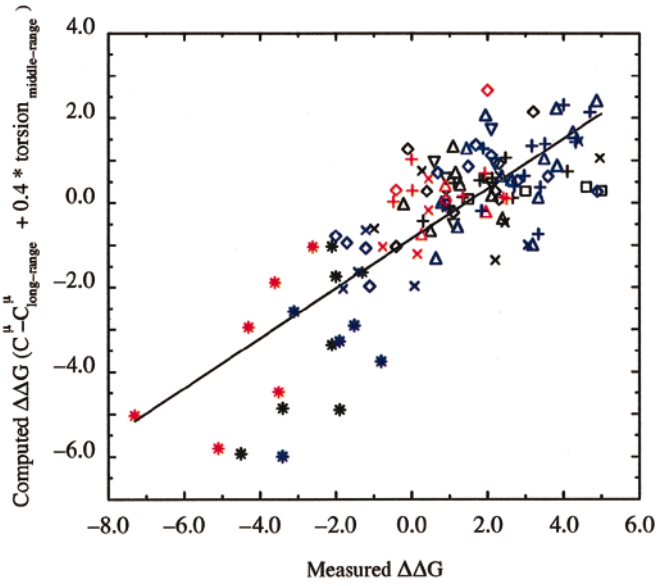

Figure 1. ΔΔ*G*s computed with the sum of the C^{μ}–$C^{\mu}_{\text{long-range}}$ potential and the $\text{torsion}_{\text{middle-range}}$ potential weighted by a factor of 0.4, as a function of the measured ΔΔ*G*s for the 121 mutations of totally buried residues listed in Table 1A. The ΔΔ*G*s are in kcal/mol. The mutations in barnase, T4 lysozyme, human lysozyme, Trp synthase, chymotrypsin inhibitor 2, chicken lysozyme and apomyoglobin are indicated with the symbols +, ◇, □, *, △, × and ▽, respectively. Mutations, for which the van der Waals radii of the mutated and mutant amino acids, defined by Levitt (1976), differ by less than 0.1 Å or by more than 0.5 Å are represented in red and blue, respectively. The line corresponds to the regression line on the 121 mutations; its equation is : $y = 0.59\ x - 0.85$. The linear correlation coefficient between measured and computed ΔΔ*G*s is equal to 0.80.

is not an artifact due to the optimization of the weighting factors. This test consists of shuffling the ΔΔ*G*s computed with the $\text{torsion}_{\text{middle-range}}$ potentials for the 121 mutations, and adding these values to the ΔΔ*G*s computed with the C^{μ}–$C^{\mu}_{\text{long-range}}$ potential, using all possible values of the weighting coefficients. It turns out that a 0.02 improvement of the correlation is observed, for some values of the weighting factors, in only one of the 1000 trials. Thus, these different tests all lead to the same conclusion: though the increase of the correlation coefficient upon adding a contribution from the torsion potential is small, it is nevertheless statistically significant.

This result means that, though the hydrophobic interactions dominate for fully buried residues, the local interactions along the sequence are not negligible. This seems *a priori* to contradict the better performance of the C^{μ}–$C^{\mu}_{\text{long-range}}$ potential compared to the C^{μ}–C^{μ} potential. However, the local interactions described by the C^{μ}–C^{μ} potential are based on propensities of residue pairs separated by a given distance along the sequence to be separated by a certain spatial distance. These are not equivalent to the local interactions described by the torsion potentials, which are based on propensities of residues to adopt certain main-chain torsion angles and are more closely related to secondary structure propensities. Thus, these seemingly contradictory results can be reconciled by stating that the local interactions along the chain, responsible for secondary structure formation, are non-negligible in the protein core.

To confirm the non-negligible effect of local interactions along the chain even for the most buried residues, we repeat the above analysis for the subset of mutations of residues with a solvent accessibility in the 0 to 5% range. It could indeed be suspected that this effect would be due to the mutations of residues with almost 20% solvent accessibility. This is not the case. We find indeed that the potential leading to the best correlation in the 0 to 5% subset is the combination of the C^{μ}–$C^{\mu}_{\text{long-range}}$ potential with the $\text{torsion}_{\text{middle-range}}$ potential weighted by a factor of 0.5. The correlation coefficient is even higher than that of the 0 to 20% set: it is equal to 0.83. The change in weighting coefficient from 0.4 to 0.5 is not significant, since its standard deviation is equal to 0.1, as mentioned above. So, the contribution of local interactions along the chain is really non-negligible in the protein core.

Mutations of Residues with a Solvent Accessibility Between 20 and 40%

For the 69 mutations of partially buried residues with solvent accessibility comprised between 20 and 40%, listed in Table1B, neither torsion potentials nor distance potentials taken individually yield good correlations between computed and experimental ΔΔ*G*s. The correlation coefficient is indeed equal to 0.57 using the best performing torsion potential, the $\text{torsion}_{\text{short-range}}$ potential, and to 0.58 using the best performing distance potential, the C^{μ}–C^{μ} potential. This is not surprising, considering that fully buried residues are dominated by hydrophobic interactions, as shown in the previous section, and that solvent accessible residues are, as for them, dominated by local interactions along the sequence, as shown by Gilis & Rooman (1996). It seems thus logical that both types of interactions are important for partially buried residues.

As expected, combinations of torsion and distance potentials perform much better than either potential taken individually. The highest correlation coefficient is obtained with the combination of the $\text{torsion}_{\text{short-range}}$ potential weighted by a factor of 1 and the C^{μ}–C^{μ} potential weighted by a factor of 0.7 (Figure 2). What emerges from this result is that the importance of local interactions along the chain gains ground when moving from the protein core towards the surface. Indeed, the weight of the torsion potential relative to the distance potential increases and moreover, the C^{μ}–C^{μ} potential, which contains a contribution

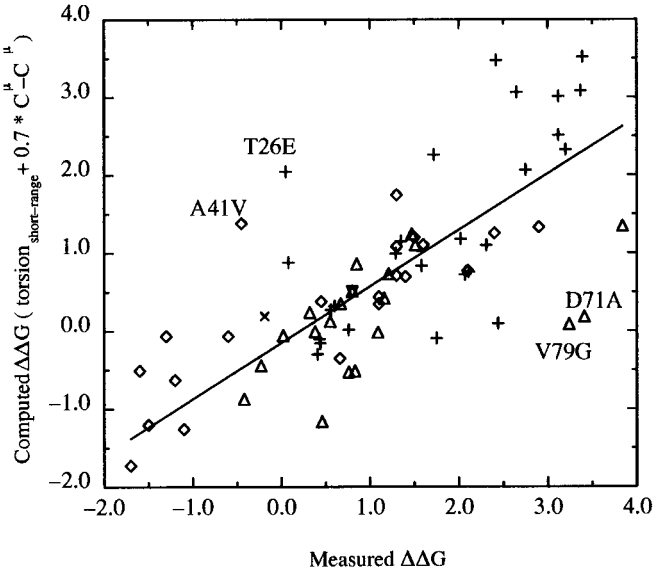

Figure 2. $\Delta\Delta G$s computed with the sum of the $\text{torsion}_{\text{short-range}}$ potential and the C^{μ}–C^{μ} potential weighted by a factor of 0.7, as a function of the measured $\Delta\Delta G$s for the 69 mutations of partially buried residues with a solvent accessibility between 20 and 40%, listed in Table 1B. The $\Delta\Delta G$s are in kcal/mol. The mutations in barnase, T4 lysozyme, chymotrypsin inhibitor 2, chicken lysozyme and apomyoglobin are indicated with the symbols +, ◇, △, × and ▽, respectively. The line corresponds to the regression line obtained with four out of the 69 mutations excluded by our sorting procedure; its equation is: $y = 0.72\ x - 0.15$. The excluded mutations are indicated by the name of the mutated amino acid, followed by their position in the sequence, followed by the name of the mutant amino acid. The linear correlation coefficient between measured and computed $\Delta\Delta G$s is equal to 0.71 on all 69 mutations, and to 0.82 when the four mutations are rejected.

from local interactions, starts to perform better than the C^{μ}–$C^{\mu}_{\text{long-range}}$ potential.

However, the correlation coefficient obtained with the combination of the $\text{torsion}_{\text{short-range}}$ potential and the C^{μ}–C^{μ} potential is not very high ($r = 0.71$), though higher than that of the individual potentials. Looking at Figure 2, we observe that the low correlation coefficient is due to a few mutations that are far from the main group. According to our automatic sorting procedure (see Methods), the four mutations that must be excluded to get better correlations are Thr26 → Glu in barnase, Ala41 → Val in T4 lysozyme and Asp71 → Ala and Val79 → Gly in chymotrypsin inhibitor 2. The correlation coefficient on the 65 remaining mutations is equal to 0.82.

Several reasons can be invoked to explain why these four mutations are far from the regression line: the structures of the native or denatured states may be modified upon mutation, thereby contradicting the basic hypothesis of our approach, or the relevant interactions are not well described by the considered combination of torsion and C^{μ}–C^{μ} potentials. It seems that the second explanation holds in the case of the four excluded mutations. Indeed, as shown in Figure 3(a), the mutation Thr26 → Glu in barnase fits well in the group of 96 mutations of solvent accessible residues considered by Gilis & Rooman (1996), with $\Delta\Delta G$s computed by the $\text{torsion}_{\text{short-range}}$ potential. The stability change caused by this mutation seems thus to be essentially governed by local interactions along the sequence. The three other excluded mutations, Ala41 → Val in T4 lysozyme, Asp71 → Ala and Val79 → Gly in chymotrypsin inhibitor 2, fit well in the group of mutations of residues that have a solvent accessibility of less than 20%, with $\Delta\Delta G$s computed with the combination of the C^{μ}–$C^{\mu}_{\text{long-range}}$ potential and the $\text{torsion}_{\text{middle-range}}$ potential weighted by a factor of 0.4, as seen in Figure 3(c). For these mutations, hydrophobic interactions seem preponderant.

There is thus a clear dependence of the best performing potentials on the solvent accessibility of the mutated residues. This dependence is, however, not absolute: the importance of local *versus* non-local interactions along the chain is not always identical for residues with the same solvent accessibility, but some fluctuations in the balance between these interactions may appear. The optimal combination of potential terms selected for each set of mutations, characterized by a certain range of solvent accessibility, corresponds to the best compromise for all mutations in the set.

Mutations of Residues with Solvent Accessibility Between 40 and 50%

The 48 mutations of residues with solvent accessibility in this range, listed in Table 1C, do not seem to present common characteristics, contrary to all other mutations. Indeed, none of the tested potentials, alone or in combination, yields reasonably high correlation coefficients: r is at most equal to 0.55. This result means either that the interactions causing the stability changes are not the same for these 48 mutations, or that some mutations perturb the structure of the backbone or denatured state. To determine if the first explanation is the right one, we investigate if subgroups of the 48 mutations fit in the three other considered sets of mutants, i.e. the sets of mutations of totally buried, partially buried and solvent accessible residues, with solvent accessibilities of less than 20%, between 20% and 40% and larger than 50%, respectively. This is done as follows. We start by identifying which of the 48 mutations fits best in the set of mutations of solvent accessible residues, that is, which mutation leads to the largest increase in the correlation coefficient. This mutation is then added to the set and the procedure is repeated until none of the remaining mutations increases the correlation coefficient (or, more precisely, does not decrease it by more than 0.02). Then, using the

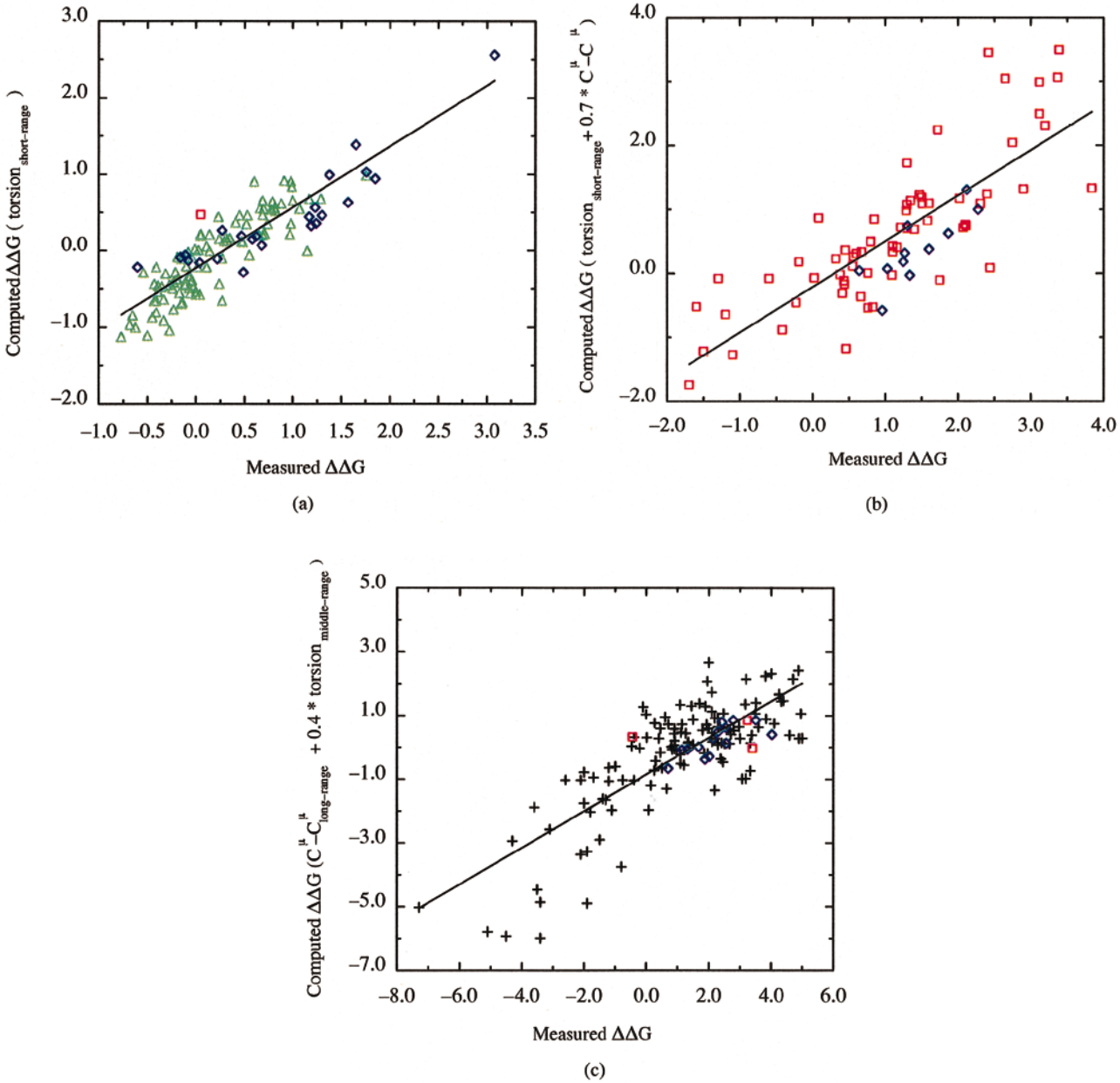

Figure 3. Computed ΔΔ*G*s as a function of measured ΔΔ*G*s, for subsets of mutations whose stability changes are mainly due to local interactions along the chain (a), non-local interactions (c) or both (b). Mutations of residues that have a solvent accessibility larger than 50%, between 40 and 50%, between 20 and 40% and smaller than 20% are indicated by the symbols △, ◇, □ and +, respectively. The ΔΔ*G*s are in kcal/mol. (a) ΔΔ*G*s computed with the torsion$_{\text{short-range}}$ potential, as a function of the measured ΔΔ*G*s, for the 120 mutations that are dominated by local interactions along the sequence. Among these 120 mutations, 96 have a solvent accessibility of at least 50% and are listed by Gilis & Rooman (1996); for 23 the accessibility is in the range 40 to 50% and for one, it is in the range 20 to 40%. The linear correlation coefficient between measured and computed ΔΔ*G*s is 0.89. The line corresponds to the regression line, whose equation is $y = 0.79x - 0.23$. (b) ΔΔ*G*s computed with the sum of the torsion$_{\text{short-range}}$ potential and the C^{μ}–C^{μ} potential weighted by a factor of 0.7, as a function of the measured ΔΔ*G*s, for the 76 mutations for which local and non-local interactions along the chain are roughly equally important. Among these, 65 have a solvent accessibility in the 20 to 40% range and 11 in the 40 to 50% range. The linear correlation coefficient between measured and computed ΔΔ*G*s is 0.80. The line corresponds to the regression line and its equation is $y = 0.71$ $x - 0.21$. (c) ΔΔ*G*s computed with the sum of the $C^{\mu} - C^{\mu}_{\text{long-range}}$ potential and the torsion$_{\text{middle-range}}$ potential weighted by a factor of 0.4, as a function of the measured ΔΔ*G*s, for the 138 mutations that are dominated by non-local interactions. Among these, 121 have a solvent accessibility of less than 20%, three have an accessibility in the 20 to 40% range and 14 in the 40 to 50% range. The linear correlation coefficient between measured and computed ΔΔ*G*s is 0.79. The line corresponds to the regression line; its equation is $y = 0.57\ x - 0.85$.

same procedure, we identify which of the remaining mutations fit in the set of mutations of partially buried residues, and finally, if the yet remaining mutations fit in the set of totally buried residues.

Following this procedure, the mutations of residues with solvent accessibility between 40 and 50% are divided into three groups. A group of 23 mutations is added to the set of 96 mutations of solvent accessible residues (Figure 3(a)), another group of 11 mutations is added to the 65 mutations of partially buried residues (Figure 3(b)), and the 14 remaining mutations are added to the set of 121 mutations of completely buried residues (Figure 3(c)). The correlation coefficients of these three sets of 119, 76 and 135 mutants are equal to 0.89, 0.80 and 0.80, respectively, and are thus quite good.

Thus, in the 40 to 50% solvent accessibility range, about half of the mutations can be included in the ensemble of mutations of solvent accessible residues and are hence dominated by local interactions along the sequence. In roughly another quarter of the mutations, local and non-local interactions are of the same order of magnitude, and in the last quarter, non-local interactions dominate. It thus appears that solvent accessibility is not a good measure for determining what the dominant interactions are in the 40 to 50% accessibility range.

Prediction Accuracy Reached with Distance and Torsion Potentials

To estimate the relative precision with which distance and torsion potentials evaluate $\Delta\Delta G$s, the plots of measured $\Delta\Delta G$s *versus* $\Delta\Delta G$s computed with either of the two potentials are compared. In particular, the plot containing the 121 mutations of totally buried residues with $\Delta\Delta G$s computed with the C^{μ}–$C^{\mu}_{\text{long-range}}$ potential is superimposed with the plot containing the 106 mutations of solvent accessible residues with $\Delta\Delta G$s computed with the torsion$_{\text{short-range}}$ potential (Figure 4). In doing so, the computed $\Delta\Delta G$s for the surface residues were rescaled in such a way that their regression line coincides with the regression line of the mutations of completely buried residues, leaving the correlation coefficient unchanged. This allows an easier comparison of the two potentials. As seen in Figure 4, the dispersion of the points around the regression line is much larger for the distance potential than for the torsion potential. Even the ten mutations of surface residues, which are considered to be far from the regression line and are excluded from the correlation, are closer to the regression line than many mutations of buried residues. Thus, the distance potential measures less well the stability changes upon mutation for totally buried residues than the torsion potential does for surface residues.

The reason for the lesser performance of distance potentials is not obvious. It can be argued that it is due to the fact that only distances between residue pairs are considered; correlations between residue triplets and quadruplets are not taken into account. This does not seem to be crucial, however. We tested indeed the triplet potential C^{μ}–C^{μ}–C^{μ}, based on contacts between three residues (see Methods). With this potential, the measured and computed $\Delta\Delta G$s have a correlation coefficient of 0.75, which is slightly lower than that obtained with the pair potential C^{μ}–C^{μ}. It is thus certainly not the neglect of triplet correlations that explains the large dispersion around the regression line. The reason why the C^{μ}–C^{μ}–C^{μ} potential gives somewhat less good results than the C^{μ}–C^{μ} poten-

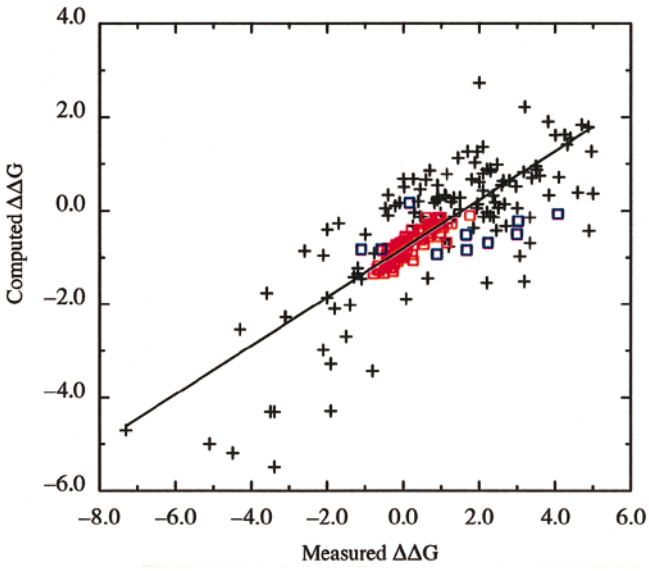

Figure 4. Dispersion of the $\Delta\Delta G$s of totally buried residues computed with a C^{μ}–C^{μ} potential compared to that of the $\Delta\Delta G$s of surface residues computed with a torsion potential. The $\Delta\Delta G$s are in kcal/mol. The mutations indicated by the symbol + correspond to the 121 mutations of totally buried residues listed in Table 1A. Their $\Delta\Delta G$s are computed with the C^{μ}–$C^{\mu}_{\text{long-range}}$ potential. The line corresponds to the regression line on these 121 mutations; its equation is: $y = 0.59\ x - 0.85$. The linear correlation coefficient between measured and computed $\Delta\Delta G$s is equal to 0.80. The mutations indicated by the symbols ▫ and ▫ correspond to the 106 mutations of surface residues, with solvent accessibility of at least 50%, analysed by Gilis & Rooman (1996). The red symbols correspond to the subset of 96 mutations for which measured and computed $\Delta\Delta G$s correlate well and the blue symbols correspond to the ten mutations that are considered to be too far from the regression line and are excluded from the correlation; these ten mutations seem to modify the native or denatured states or imply interactions that are atypical for surface residues. The $\Delta\Delta G$s of these 106 mutations are computed with the torsion$_{\text{short-range}}$ potential. The regression line is obtained on the 96 mutations and the corresponding correlation coefficient is equal to 0.87. The $\Delta\Delta G$s on the 106 mutations of surface residues have been rescaled in such a way that their regression line coincides with the regression line of the 121 mutations of buried residues, without changing the correlation coefficient. This allows us to compare the dispersion around the regression line on both sets.

tial may be attributed to the fact that the former is a contact-non-contact potential whereas the latter is a more precise distance-dependent potential, and also, that the statistics on residue triplets are less reliable than on residue pairs because of the limited data.

Another reason for the limited performance of distance potentials on buried residues might be that the backbone structure is modified upon mutation. This hypothesis can readily be tested on eight out of the 121 mutations of completely buried residues, whose structures have been solved by X-ray crystallography to better than 2 Å resolution. They correspond to the mutations Ile3 → Tyr,Gly77 → Ala,Ala98 → Val,Leu133 → Ala,Asn149 → Cys and Thr152 → Ser in T4 lysozyme and Tyr78 → Phe and Ser91 → Ala in barnase. For these eight mutants, the $\Delta\Delta G$ can be computed as the difference between the folding free energy $\Delta G^S(C_m)$ of the mutant structure C_m and the folding free energy $\Delta G^S(C_w)$ of the wild-type structure C_w, or, as in the previous sections, with the hypothesis that the mutant and wild-type backbone structures coincide, that is, $C_m = C_w$. The $\Delta\Delta G$s computed in these two ways turn out to be similar. Indeed, the correlation coefficient on the 121 mutations is equal to 0.80, whether the eight above mentioned mutations are computed with the $C_m = C_w$ assumption or not. Hence, the modification of the backbone structure does not seem, at least from this result, to be responsible for the large dispersion of $\Delta\Delta G$ values around the regression line, observed for mutations of buried residues.

This result justifies furthermore the approximation made throughout this study that wild-type and mutant proteins have the same backbone structure ($C_m = C_w$). This approximation turns out to be even more accurate for mutations of buried residues than for mutations of surface residues, where some of the mutations had to be excluded from the correlation because they caused structural rearrangements implying the modification of backbone torsion angle domains. This can easily be understood, considering that residues in the protein core are much more constrained and only small backbone rearrangements can occur, which are hardly detectable by our potentials. Moreover, even if the structure of the wild-type and mutant proteins are both available, it is sometimes better to overlook one of the structures and to use the $C_m = C_w$ approximation, especially when one of the structures is not well resolved. For example, for the mutation Ile3 → Leu in T4 lysozyme, both the wild-type and mutant structures are known, but the latter has a resolution of 2.6 Å and the root mean square deviation of the backbone coordinates after superposition is as high as 1.7 Å; the structural differences are not confined in the region of the mutation, but are dispersed over the whole structure. As a result, the computed $\Delta\Delta G$s differ by 1.9 kcal/mol, according to whether the approximation $C_m = C_w$ is used or not. This difference is non-physical, it is only due to the limited resolution of one of the structures.

Finally, a last reason that could explain the lesser performance of the C^μ–C^μ potential compared to the torsion potential is that the former does not describe sufficiently accurately the interactions that stabilize the residues in the protein core. It has been shown that an important factor of stability changes upon mutation of residues in the protein core is the cavity formation when mutating a large into a small amino acid and the strain caused by mutating a small into a large amino acid (Kellis *et al*., 1989; Eriksson *et al*., 1992; Kocher *et al*., 1996). The stability changes differ according to the environment of the mutated residues. If there is residual flexibility in the structure, cavities can be more readily filled and strain can be relaxed. It is not obvious whether database-derived potentials account for this effect. Torsion potentials certainly do not, given that they represent local interactions along the chain. Distance potentials could in principle account for it, as they describe the spatial environment of residues. However, it is not sure that they are sensitive enough, as the created and filled cavities are small compared to the distance precision of the potentials. This is more especially true as distances are computed between average side-chain centroids, which can occasionally be rather different from the exact ones and have steric overlaps that do not occur in the true structures. Moreover, due to the approximation that the environment can be described by pair or triplet interactions, a global view of the environment of a given residue, necessary for detecting cavities, is lacking.

In order to support the hypothesis that the width of the distribution around the regression line can at least in part be explained by the neglect of effects due to cavity formation or filling, we identify the mutations of totally buried residues, among the 121 considered ones, for which the mutated and mutant amino acids have roughly the same size. Two amino acids are defined as having similar sizes if their radii, defined by Levitt (1976), differ by 0.1 Å at most. This subgroup contains the 23 mutations depicted in red in Figure 1. When correlating their computed $\Delta\Delta G$s with their measured $\Delta\Delta G$s, we find a correlation coefficient of 0.87, against 0.80 for the whole set of 121 mutations. On the contrary, the correlation coefficient of the 50 mutations where the mutated and mutant amino acids differ significantly in size, as measured by a change in the amino acid radius of more than 0.5 Å, depicted in Figure 1 in blue, is equal to 0.80 and thus identical to that computed on the whole set.

To assess the statistical significance of the increase of the correlation coefficient from 0.80 for the full set of 121 mutations up to 0.87 for the subset of 23 mutations with almost no size modification, we randomly choose 23 mutations out of the full set and compute the correlation coefficient r on this subset. Repeating this procedure 1000

times, it is found that r is equal to 0.87 or more in 10% of the subsets. At first sight, there seems thus to be 90% chance that the observed increase is statistically significant. But one has to remember that the randomly picked mutations usually do not have any common property, contrary to the 23 mutations chosen on the basis of residue size. The chance of picking 23 well correlating mutations sharing a common property is difficult to estimate, but is obviously much lower than 10%. Thus, we cannot be absolutely sure that the increase of the correlation coefficient from 0.80 up to 0.87 is meaningful, but the confidence level is certainly much higher than 90%.

We can hence conclude that the correlation between measured $\Delta\Delta G$s and $\Delta\Delta G$s computed with the C^{μ}–C^{μ} potential for completely buried residues appears to be significantly higher when the sizes of the mutant and mutated residues are similar and no cavities are created or filled. This result does not prove, but brings clear support to the hypothesis that the factor limiting the ability of distance potentials to evaluate the $\Delta\Delta G$s of buried residues is that they do not take properly into account the effect of cavities in the protein core.

Conclusion

It was shown here that the crucial factor determining the relative importance of local *versus* non-local interactions along the chain is the solvent accessibility of the mutated residues. The dominant interactions for surface residues, with solvent accessibilities of at least 50%, are local along the chain and well described by torsion potentials, whereas the dominant interactions for totally buried residues, with solvent accessibilities of 20% at most, are non-local along the chain and well described by the C^{μ}–C^{μ} potentials dominated by hydrophobic forces. However, we would like emphasize that the local interactions responsible for secondary structure formation, though less important than hydrophobic interactions, play a non-negligible role even for the most buried residues. The importance of these interactions is often unduly overlooked. For partially buried residues, with solvent accessibilities between 20 and 50%, local and non-local interactions are, on the average, roughly equally important. Their relative importance can vary from one position to another, depending on the type of mutated and mutant amino acid and of their environment. This variation is especially strong for mutated residues with accessibilities between 40 and 50%. In this twilight zone, the solvent accessibility of the mutated residue is not a good measure for determining what the dominant interactions are. If it would be possible to know the solvent accessibility of both the mutated and mutant residues, one could probably estimate more reliably the relative importance of local and non-local interactions. But this information is usually not available, as only a few mutant structures have been determined. It could be envisaged to position the side-chains of the mutant sequence in the wild-type structure using side-chain positioning algorithms, and to compute the solvent accessibility of the mutant residue in the so-modelled structure. But the drawback of this approach is that side-chain positioning algorithms are not 100% reliable.

The local and non-local interactions along the chain described by the torsion and C^{μ}–C^{μ} potentials, respectively, are not the only important interactions that contribute to stabilize the protein core. Another important (de)stabilizing effect is due to cavity formation or filling. This effect is particularly important for certain mutations, involving residues in closely packed environments with no residual flexibility. Our analysis seems to indicate that the potentials used fail to correctly describe this effect. In principle, it would be possible to define a new kind of database-derived potential that would be specifically designed to account for it. It is, however, not obvious that such a potential can be constructed in practice. Indeed, this effect requires a description at the atomic level of detail, where small cavities and strain provoked by somewhat too closely packed atoms come into play. Such effects are difficult to account for using residue-based effective potentials. But if one succeeds in designing a potential that accounts for these effects in a satisfactory fashion, it would be a great achievement, as it seems to be the last important type of interaction not represented by any database-derived potential. Such new potentials contain the promise of improving significantly all structure prediction algorithms, whether it is fold recognition, inverse folding, *ab initio* predictions or prediction of stability changes upon mutation.

Though the correlation between computed and measured $\Delta\Delta G$s is not perfect, it is far from bad and can be used for prediction purposes to yield a first estimation of the stability changes to be expected. We would like to stress that the predictive value of our procedure owes to the fact that the correlations are valid for mutations at different sites and in different proteins. For about 90% of the mutated residues with a solvent accessibility of more than 50%, predicted and measured $\Delta\Delta G$s have a correlation coefficient of 0.87. Among the mutations of residues whose accessibility is in the 20 to 40% range, 95% have a correlation coefficient of 0.82. And for the ensemble of mutated residues with an accessibility of at most 20%, the correlation coefficient is of 0.80; this coefficient increases up to 0.83 for the subset of mutated residues with at most 5% solvent accessibility. Only for the mutations of residues whose solvent accessibility is in the 40 to 50% range, does the predictive power of our procedure break down. But yet, the $\Delta\Delta G$s of these mutations can be estimated in three different ways, by using the optimal potential of the set of totally buried, partially buried or surface residues. There remains, however, an uncertainty about

which of the three computed $\Delta\Delta G$s must be trusted.

Methods

Formalism for deriving effective potentials from known structures

To derive potentials from known protein structures, sequences S are divided into sequence elements s (e.g. residues, residue pairs or residue triplets) and conformations C are divided into structural states c (e.g. ranges of torsion angles or inter-residue distances). The frequencies of c and s in the dataset of known proteins are computed, yielding an estimation of the probability of c, $P(c)$, and the probability of c knowing s, $P(c|s)$. Using the formalism described by Rooman & Wodak (1995), these probabilities are related to the folding free energy, $\Delta G^S(C)$, using:

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

where the indices i and j indicate the position(s) along the sequence of the structural states and sequence elements, respectively, k is the Boltzmann constant and T a conformational temperature (Pohl, 1971) taken to be room temperature. The folding free energy, $\Delta G^S(C)$, is the difference between the free energy of the sequence S adopting a conformation C,$G^S(C)$, and the free energy of a denatured-like state of S,G^S, in which the conformational states c and the sequence elements s are uncorrelated.

Torsion potentials

Torsion potentials are computed from the propensities of residues to be associated to 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).

Two types of torsion potentials have been previously developed and tested (Rooman *et al.*, 1991, 1992; Kocher *et al.*, 1994; Gilis & Rooman, 1996). Only the most performing one is considered here. It takes into account 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 (1) becomes:

$$\Delta G^S(C) = -kT \sum_{i,j,k=1}^{N} \frac{1}{\zeta_k} \log \frac{P(t_i, t_j|a_k)}{P(t_i, t_j)} \quad (2)$$

where N is the number of residues in the sequence S. We consider a ''short range'' backbone potential, noted $\text{torsion}_{\text{short-range}}$, 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, $\text{torsion}_{\text{middle-range}}$, 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.

Distance potentials

Pair potentials

Pair potentials 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). Probabilities of residues separated by one to six positions along the sequence are computed separately, whereas probabilities of residues separated by seven positions and more are all merged. This distinction yields potentials that represent both local and non-local interactions along the chain. The folding free energy defined by these potentials is, according to equation (1):

$$\Delta G^S(C) = -kT \sum_{i,j=1}^{N} \log \frac{P^{|i-j|}(d_{ij}|a_i, a_j)}{P^{|i-j|}(d_{ij})} \quad (3)$$

with $i+1 < j$ and with the probabilities $P^{|i-j|}$ being independent of $|i-j|$ for $|i-j| > 7$. The inter-residue distances d_{ij} are computed between the average centroids, C^μ, which are specific to each amino acid type and are defined as the average of the atomic coordinate centres of all conformations of side-chains of the same type observed in the protein dataset (Kocher *et al.*, 1994). The distances between 3 and 8 Å are grouped into 25 bins of 0.2 Å width and the distances larger than 8 Å are merged. Further details are given in Kocher *et al.* (1994). This potential, called C^μ–C^μ potential, yields better performances than the C^α–C^α and C^β–C^β potentials, where the inter-residue distances are computed between C^αs and C^βs, respectively. Thus, only results with the former are presented.

In a variant of this potential, only residues separated by more than 15 residues along the sequence ($|i-j| > 15$) are taken into account. The value of 15 has been obtained by optimizing the correlations between measured and computed $\Delta\Delta G$s on the set of mutations of completely buried residues. This potential represents only non-local interactions along the sequence and is referred to as C^μ–$C^\mu_{\text{long-range}}$ potential.

Triplet potentials

Triplet potentials are computed from propensities of three residues a_i, a_j, a_k to be in contact. Two residues a_i and a_j are considered to be in contact when their spatial distance d_{ij} is less than 7 Å. To have valid statistics, only contact and non-contact bins are considered: one in which all three residues

are simultaneously in contact, one in which none of the three are in contact, and six where there are contacts between residues, but not simultaneously between the three residues. Residues separated by less than 11 positions along the chain are not taken into account; this value has been adjusted to get the best correlation between computed and measured $\Delta\Delta G$s on the set of mutations of completely buried mutations. According to equation (1), the folding free energy is:

$$\Delta G^S(C) = -kT \sum_{i,j=1}^{N} \log \frac{P(d_{ij}, d_{ik}, d_{jk}|a_i, a_j, a_k)}{P(d_{ij}, d_{ik}, d_{jk})} \quad (4)$$

The inter-residues distances d_{ij}, d_{ik}, d_{jk} are computed between the average centroids $C^\mu$and the potential is called C^μ–C^μ–C^μ potential.

Correction for sparse data

When computing all the above mentioned potentials, the correction for sparse data described by Gilis & Rooman (1996) is applied.

Weighted combination of potential terms

Linear combinations of different potential terms are tested. In particular, we consider the combinations of the type $A\Delta G^S(C)^{(1)} + B\Delta G^S(C)^{(2)}$ where $\Delta G^S(C)^{(1)}$ corresponds to the $\text{torsion}_{\text{short-range}}$ or $\text{torsion}_{\text{middle-range}}$ potentials and $\Delta G^S(C)^{(2)}$ to the C^μ–C^μ or C^μ–$C^\mu_{\text{long-range}}$ potentials. A and B are real values between 0 and 1; the tested values are A, $B = 0.0, 0.1, 0.2, \ldots, 1.0$.

Protein structure data

The potentials are derived from a set of 141 well resolved ($\leqslant$2.5 Å) and refined proteins from the Brookhaven databank (Bernstein *et al*., 1997), with less than 20% sequence identity or no structural homology. A list of these proteins can be found in Wintjens *et al*. (1996). To avoid biasing the predictions towards the native structure, the potentials are derived from all proteins from the set except those that have more than 20% sequence identity with the protein on which predictions are performed.

Computing folding free energy changes upon mutation

To compare our results with experimental ones, we have to calculate a difference in folding free energy, $\Delta\Delta G$, between mutant and wild-type. In a first step, we compute the folding free energy of the mutant protein, $\Delta G^{mutant}(C_w)$, and of the wild-type protein, $\Delta G^{wild\text{-}type}(C_w)$, using equations (2) to (4). C_w corresponds to the native structure of the wild-type. It is thus assumed, unless stated otherwise, that the wild-type and mutant proteins have the same backbone structure. The $\Delta\Delta G$ is calculated using the sign convention:

$$\Delta\Delta G = \Delta G^{mutant}(C_w) - \Delta G^{wild\text{-}type}(C_w) \quad (5)$$

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

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

To compare experimentally determined $\Delta\Delta G$s and computed $\Delta\Delta G$s of a set of mutations, a correlation coefficient is calculated, assuming a linear regression. To estimate the significance of this correlation, the probability $\mathscr{P}$ that the same correlation would arise by random sampling in an uncorrelated population is computed (Fisher, 1958). For all sets of mutations considered in this paper, $\mathscr{P}$ is lower than 10^{-6}; the correlations are thus undoubtedly statistically significant.

Another issue is the error on the computed correlation coefficients. To estimate this error, we take the full set of mutations, drop a number of randomly chosen mutations and compute the correlation coefficient on the remaining mutations. This procedure is repeated 1000 times. The average correlation coefficient on these 1000 trials and the standard deviation is then computed, thereby giving an estimation of the validity of the computed correlation coefficients.

Automatic sorting procedure

An automatic sorting procedure is used to determine which of the mutations in a given set are responsible for the low value of the correlation coefficient. It proceeds by rejecting the mutation that leads to the highest correlation coefficient for the mutations remaining in the set. This procedure is repeated until the correlation coefficient exceeds a certain value.

Acknowledgements

We thank Jean-Pierre Kocher and Martine Prévost 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 Belgium National Fund for Scientific Research (FNRS).

References

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

Alber, T., Daopin, S., Wilson, K., Wozniak, J. A., Cook, S. P. & Matthews, B. W. (1987). Contributions of hydrogen bonds of Thr157 to the thermodynamic stability of phage T4 lysozyme. *Nature*, **330**, 41–46.

Basch, P. A., Singh, U. C., Landridge, 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 archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–542.

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

Casari, G. & Sippl, M. (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.

Daopin, S., Baase, W. A. & Matthews, B. W. (1990). A mutant T4 lysozyme (Val 131 → Ala) designed to increase thermostability by the reduction of strain within an α-helix. *Proteins: Struct. Funct. Genet.* **7**, 198–204.

Daopin, S., Alber, T., Baase, W. A., Wozniak, J. A. & Matthews, B. W. (1991). Structural and thermodynamic analysis of the packing of two α-helices in bacteriophage T4 lysozyme. *J. Mol. Biol.* **221**, 647–667.

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. (1995). Optimization of rates of protein folding: the nucleation-condensation mechanism and its implications. *Proc. Natl Acad. Sci. USA,* **92**, 1086–10873.

Fersht, A. R., Matouschek, A. & Serrano, L. (1992). The folding of an enzyme. I. Theory of protein engineering analysis of stability and pathways of protein folding. *J. Mol. Biol.* **224**, 771–782.

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

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.

Gō, N. & Taketomi, H. (1978). Respective roles of short- and long-range interactions in protein folding. *Proc. Natl Acad. Sci. USA,* **75**, 559–563.

Govindarajan, S. & Goldstein, R. (1995). Optimal local propensities for model proteins. *Proteins: Struct. Funct. Genet.* **22**, 413–418.

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 chymotrypsin inhibitor 2 to protein stability and activity. *Biochemistry,* **33**, 13880–13887.

Jackson, S. E., Moracci, M., elMasry, N., Johnson, C. M. & Fersht, A. R. (1993). Effect of cavity-creating mutations in the hydrophobic core of chymotrypsin inhinbitor 2. *Biochemistry,* **32**, 11259–11269.

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

Kay, M. S. & Baldwin, R. L. (1996). Packing interactions in the apomyoglobin folding intermediate. *Nature Struct. Biol.* **3**, 439–445.

Kellis, J. T., Jr, Nyberg, K., Sali, D. & Fersht, A. R. (1988). Contribution of hydrophobic interactions to protein stability. *Nature,* **333**, 784–786.

Kellis, J. T., 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.

Kocher, J.-P., Prévost, M., Wodak, S. J. & Lee, B. (1996). Properties of the protein matrix revealed by the free energy of cavity formation. *Structure,* **4**, 1517–1529.

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.

Levitt, M. (1976). A simplified representation of protein conformations for rapid simulation of protein folding. *J. Mol. Biol.* **104**, 59–107.

Matouschek, 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.

Matsumura, M., Becktel, W. J. & Matthews, B. W. (1988). Hydrophobic stabilization in T4 lysozyme determined directly by multiple substitutions of Ile 3. *Nature,* **334**, 406–410.

Matthews, B. W., Nicholson, H. & Becktel, W. J. (1987). Enhanced protein thermostability from site-directed mutations that decrease the entropy of unfolding. *Proc. Natl Acad. Sci. USA,* **84**, 6663–6667.

Matthews, S. J., Jandu, S. K. & Leatherbarrow, R. J. (1993). ^{13}C NMR study of the effects of mutation on the tryptophan dynamics in chymotrypsin inhibitor 2: correlations with structure and stability. *Biochemistry,* **32**, 657–662.

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 ϕ-ψ matrices: comparison with experimental data. *Proteins: Struct. Funct. Genet.* **20**, 301–311.

Muñoz, V. & Serrano, L. (1996). Local versus nonlocal interactions in protein folding and stability—an experimentalist's point of view. *Folding Design,* **1**, R71–R77.

Otzen, D. E. & Fersht, A. R. (1995). Side-chain determinants of β-sheet stability. *Biochemistry,* **34**, 5718–5724.

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

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

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.
Serrano, L. & Fersht, A. R. (1989). Capping and α-helix stability. *Nature,* **342**, 296–299.
Serrano, L., Kellis, J. T., Jr, Cann, P., Matouschek, A. & Fersht, A. R. (1992a). The folding of an enzyme. II. Substructure of barnase and the contribution of different interactions to protein stability. *J. Mol. Biol.* **224**, 783–804.
Serrano, L., Sancho, J., Hirshberg, M. & Fersht, A. R. (1992b). α-Helix stability in proteins. I. Empirical correlations concerning substitution of side-chains at the N and C-cap and the replacement of alanine by glycine or serine at solvent-exposed surfaces. *J. Mol. Biol.* **227**, 544–559.
Shakhnovich, E., Abkevich, V. & Ptitsyn, O. (1996). Conserved residues and the mechanism of protein folding. *Nature,* **379**, 96–98.
Shih, P. & Kirsch, J. F. (1995). Design and structural analysis of an engineered thermostable chicken lysozyme. *Protein Sci.* **4**, 2063–2072.
Shih, P., Holland, D. B. & Kirsch, J. F. (1995). Thermal stability determinants of chicken egg-white lysozyme core mutants: hydrophobicity, packing volume, and conserved buried water molecules. *Protein Sci.* **4**, 2050–2062.
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.
Sippl, M. J. (1995). Knowledge-based potentials for proteins. *Curr. Opin. Struct. Biol.* **5**, 229–235.
Takano, K., Ogasahara, K., Kaneda, H., Yamagata, Y., Fujii, S., Kanaya, E., Kikuchi, M., Oobatake, M. & Yutani, K. (1995). Contribution of hydrophobic residues to the stability of human lysozyme: calorimetric studies and X-ray structural analysis of the five isoleucine to valine mutants. *J. Mol. Biol.* **254**, 62–76.
Tidor, B. & Karplus, M. (1991). Simulation analysis of the stability mutant R96H of T4 lysozyme. *Biochemistry,* **30**, 3217–3228.
Unger, R. & Moult, J. (1996). Local interactions dominate folding in a simple protein model. *J. Mol. Biol.* **259**, 988–994.
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.
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 subunit. *Proc. Natl Acad. Sci. USA,* **84**, 4441–4444.
Zhang, X.-J., Baase, W. A. & Matthews, B. W. (1992). Multiple alanine replacements within α-helix 126-134 of T4 lysozyme have independent, additive effects on both structure and stability. *Protein Sci.* **1**, 761–776.
Zhang, X.-J., Baase, W. A., Shoichet, B. K., Wilson, K. P. & Matthews, B. W. (1995). Enhancement of protein stability by the combination of point mutations in T4 lysozyme is additive. *Protein Eng.* **8**, 1017–1022.

Edited by J. Thornton

(Received 12 February 1997; received in revised form 18 May 1997; accepted 25 June 1997)