Theor Chem Acc (1999) 101:46–50
DOI 10.1007/s00214980m121

Theoretical
Chemistry Accounts

Regular article

Prediction of stability changes upon single-site mutations using database-derived potentials*

Dimitri Gilis, Marianne Rooman

Ingénierie Biomoléculaire, Service de Chimie Organique, CP 165, avenue Roosevelt 50, B-1050 Brussels, Belgium

Received: 20 May 1998 / Accepted: 3 September 1998 / Published online: 7 December 1998

Abstract. One of the purposes of studying protein stability changes upon mutations is to get information about the dominating interactions that drive folding and stabilise the native structure. With this in mind, we present a method that predicts folding free-energy variations caused by point mutations using combinations of two types of database-derived potentials, i.e. backbone torsion-angle potentials and distance potentials, describing local and non-local interactions along the chain, respectively. The method is applied to evaluate the folding free-energy changes of 344 single-site mutations introduced in six different proteins and a synthetic peptide. We found that the relative importance of local versus non-local interactions along the chain is essentially a function of the solvent accessibility of the mutated residues. For the subset of totally buried residues, the optimal potential is the sum of a distance potential and a torsion potential weighted by a factor of 0.4. This combination yields a correlation coefficient between measured and computed changes in folding free energy of 0.80. For mutations of partially buried residues, the best potential is the sum of a torsion potential and a distance potential weighted by 0.7. For fully accessible residues, the torsion potentials taken alone perform best, reaching correlation coefficients of 0.87 on all but 10 mutations; the excluded mutations seem to modify the backbone structure or to involve interactions that are atypical for the surface. These results show that the relative weight of non-local interactions along the sequence decreases as the solvent accessibility of the mutated residue increases, and vanishes at the protein surface. On the contrary, the weight of local interactions increases with solvent accessibility. The latter interactions are nevertheless never negligible, even for the most buried residues.

Key words: Single-site mutations – Folding free-energies – Protein stability – Mean force potentials

1 Introduction

Protein engineering experiments have proven to be powerful tools for studying stability changes upon mutations, both in the folded and in the transition states [1]. Specific mutations are introduced via site-directed mutagenesis, and the resulting changes in unfolding free energy and in the free energy of activation are measured [2, 3]. By appropriate choice of the mutations, the stability changes can be related to the formation or breaking of specific interactions in the native structure or in folding intermediates.

Beside the experimental techniques, several theoretical approaches [4, 5] exist, their purpose being to rationalise experimental data and to predict the effect of new mutations so as to limit the number of experimental tests. Up to now, however, none of these methods has been completely satisfactory: either they use detailed atomic models [4] and are thus so computer time-consuming that they can only be applied to a few mutations, or they are based on rougher protein descriptions [5], but then the computed stability changes of mutations at different sites and in different proteins are usually not comparable, thereby limiting their predictive value.

The approach described in this paper [6–8] does not present this shortcoming: it allows the changes in folding free energy caused by mutations introduced in different sites of various proteins to be predicted with satisfactory accuracy. The feature that explains the good performance of our method is that we use linear combinations of different potentials, describing different kinds of interactions, whose coefficients essentially depend on the solvent accessibility of the mutated residues.

* Contribution to the Proceedings of Computational Chemistry and the Living World, April 20–24, 1998, Chambery, France

Correspondence to: D. Gilis

2 Set of mutations considered

The present approach is based on the assumption that the backbone conformations of the native and denatured states are only slightly affected upon mutation. It is therefore restricted to single-site mutations, which are, in general, the most liable to satisfy this assumption. In total, 344 mutations are considered, the stability changes of which have been measured experimentally [2, 3]. They are introduced at different sites and secondary structures in barnase, human, chicken and T4 lysozyme, chymotrypsin inhibitor 2, tryptophan synthase, apomyoglobin and a synthetic helical peptide. These mutations are divided into four subsets, according to the value of the solvent accessibility of the mutated residues computed by the program SurVol [9]: 106 mutations with a solvent accessibility of more than 50% (listed in Ref. [6]), 48 between 40 and 50%, 69 between 20 and 40% and 121 of less than 20% (listed in Ref. [7]).

3 Estimation of the stability changes upon mutation

To estimate the stability changes upon mutation, we compute the folding free energy ΔG of the wild-type and mutant structures, denoted C_{w} and C_{m} respectively. Assuming that point mutations only slightly modify the native backbone structure ($C_{\mathrm{w}} \approx C_{\mathrm{m}}$) and the denatured state, the folding free-energy changes are evaluated by:

$$\Delta\Delta G = \Delta G_{\mathrm{m}}(C_{\mathrm{w}}) - \Delta G_{\mathrm{w}}(C_{\mathrm{w}}) \ . \tag{1}$$

The folding free energies ΔG are estimated using mean force potentials derived from a set of 141 well-resolved and refined protein structures, with low sequence homology (see Ref. [10] for a list). Two main types of potentials are used: backbone torsion potentials and distance potentials.

3.1 Torsion potentials

Torsion potentials [11] are based on a local representation of the backbone structure in terms of seven domains in (ϕ,ψ,ω) backbone torsion angles, and describe local interactions along the chain. They are computed from propensities of amino acids a_k, at position k along the sequence, to be associated with a (ϕ,ψ,ω) domain t_i at position i along the sequence, or with pairs of (ϕ,ψ,ω) domains (t_i,t_j) at positions i and j:

$$\Delta G_s(C) = -k_B T \sum_{i,j,k=1}^{N} \frac{1}{\zeta_k} \log \frac{P(t_i, t_j \mid a_k)}{P(t_i, t_j)} \ , \tag{2}$$

where N is the number of residues in the sequence, k_B is the Boltzmann constant and T is a temperature taken to be room temperature. Two types of torsion potentials are defined. The first, called torsion$_{\text{short-range}}$, contains contributions from the residues in the sequence window $k-1 \le i \le j \le k+1$, and the second, called torsion$_{\text{middle-range}}$, from the interval $k-8 \le i \le j \le k+8$. ζ_k is a normalisation coefficient ensuring that each residue in the window is counted once; it is equal to the window width, except near chain ends.

3.2 Distance potentials

Distance potentials are based on a representation of the backbone structure in terms of spatial distances between the residues, measured here between average side chain centroids depending on the amino acid type, denoted C^{μ}. They are derived from the propensities of amino acid pairs (a_i, a_j) at positions i and j along the chain to be separated by a given spatial distance d_{ij}. We consider three different distance potentials. The first [7], denoted C^{μ}-$C^{\mu}_{\text{long-range}}$, describes exclusively non-local interactions along the chain:

$$\Delta G_s(C) = -k_B T \sum_{i<j}^{N} \log \frac{P(d_{ij}|a_i, a_j)}{P(d_{ij})} \ , \tag{3}$$

where $j > i + 15$. The second potential [12], denoted simply $C^{\mu} - C^{\mu}$, describes both local and non-local interactions. For residue pairs that are close along the chain, i.e. $1 < j - i < 8$, it is computed separately for each residue separation $j - i$:

$$\Delta G_s(C) = -k_B T \sum_{i<j}^{N} \log \frac{P^{j-i}(d_{ij}|a_i, a_j)}{P^{j-i}(d_{ij})} \ , \tag{4}$$

whereas for residues that are distant along the sequence, $j - i \ge 8$, all sequence separations are merged and the potential is given by Eq. (3). These two potentials are dominated by hydrophobic interactions. The third potential [8], referred to as C^{μ}-C^{μ}_{elec}, is quite different:

$$\Delta G_s(C) = -k_B T \sum_{i<j}^{N} \log \times \frac{P^{j-i}(d_{ij}, a_i, a_j)}{[P^{j-i}(a_i, a_j) P^{j-i}(d_{ij}, a_i) P^{j-i}(d_{ij}, a_j)] / [P^{j-i}(a_i) P^{j-i}(a_j) P^{j-i}(d_{ij})]} \ , \tag{5}$$

where a different potential is computed for each value of $j - i$ with $1 < j - i < 8$, and a single potential for all $j - i \ge 8$. Here, instead of comparing the propensity of two residues to be at a given distance with the propensity of any two residues to be at that distance, as in Eqs. (3) and (4), it is compared with the propensity of each of the two residues to be at that distance from any other residue. This potential corresponds to a different way of extracting the physical correlations between the amino acids and the tertiary structure from the bulk interactions due to the presence of many residues in a protein. It attaches less weight to hydrophobic interactions than the potentials defined by Eqs. (3) and (4). It has proven [8] to be better suited for describing proteins in an apolar medium or protein regions stabilised by electrostatic interactions.

4 Results

Linear combinations of the torsion (Eq. 2) and distance potentials (Eqs. 3–5) are used to evaluate the $\Delta\Delta G$s of the set of 344 mutations. These $\Delta\Delta G$s are then correlated to the experimentally measured ones, assuming a linear regression. We found that none of the potentials tested,

neither alone nor in combination, gives good estimation of the $\Delta\Delta G$s for the whole set of mutations. It turns out that, to reach a satisfactory accuracy level, the set must be divided into subsets depending on the solvent accessibility of the mutated residues.

4.1 Mutations of surface residues (solvent accessibility ≥ 50%)

For the 106 mutations of solvent-accessible residues, the potentials yielding the best correlations between measured and computed $\Delta\Delta G$s are the $\text{torsion}_{\text{short-range}}$ or $\text{torsion}_{\text{middle-range}}$ potentials taken alone [6]. The correlation coefficient is rather low, however: 0.67. But, as seen in Fig. 1, this low score is due to a few mutations that are far from the regression line. To identify objectively these outsiders, we use an automatic sorting procedure, which excludes one mutation at a time from the original set, until the correlation coefficient exceeds a given value; the rejected mutation is the mutation that, when discarded, gives rise to the highest correlation coefficient on the remaining mutations.

With this sorting procedure, we find that dropping only 10 out of the 106 mutations increases the correlation coefficient up to 0.87 for both the $\text{torsion}_{\text{short-range}}$ and the $\text{torsion}_{\text{middle-range}}$ potentials. The torsion potentials thus reliably estimate the stability changes of the remaining 96 mutations, and the ten excluded mutations may be suspected to possess unusual characteristics. This is indeed the case: seven of them have a proline either as mutant or as mutated amino acid and therefore very probably cause (certainly when it occurs in a helix) backbone rearrangements with switching of (ϕ,ψ,ω) domains. That this actually happens could be verified for the A82 → P mutation in T4 lysozyme, where both the wild-type and mutant structures have been determined. The departure of T105 → V in barnase from the regression line is less clear, but could perhaps also be attributed to modifications of the backbone structure.

The reason why the two mutations T16 → S and T16 → G in barnase are not well predicted is different. They involve the breaking of strong hydrophobic interactions, which are atypical for surface residues and are better evaluated by distance potentials than by torsion potentials. Hence, though T16 has a solvent accessibility of more than 50%, it fits in the set of totally buried residues (Fig. 2). The mutation D8 → A, which is only rejected by the $\text{torsion}_{\text{short-range}}$ potential, also involves interactions not well-represented by torsion potentials, i.e. electrostatic interactions.

4.2 Mutations of totally buried residues (solvent accessibility < 20%)

The potential that performs best on the 121 mutations of totally buried residues is the C^{μ}-$C^{\mu}_{\text{long-range}}$ potential, with

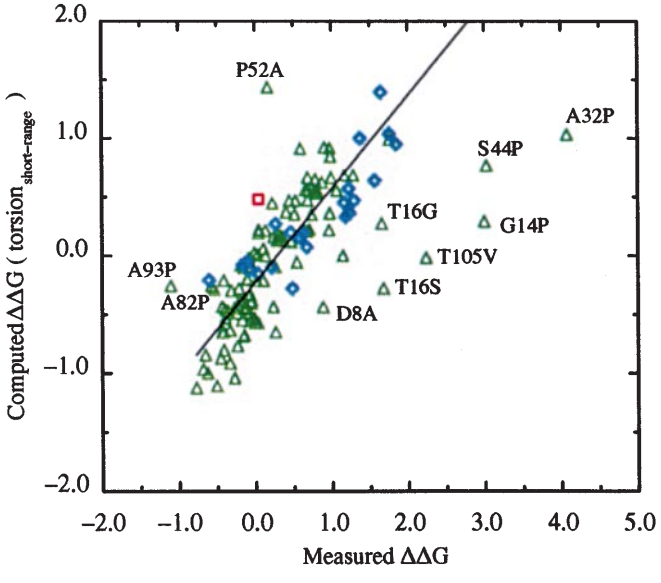

Fig. 1. $\Delta\Delta G$s computed with the $\text{torsion}_{\text{short-range}}$ potential as a function of measured $\Delta\Delta G$ (in kcal/mol) for the set of 106 surface mutations (green Δ symbols), with in addition the subset of more buried mutations that involve mainly local interactions along the chain, i.e. 23 mutations of partially exposed mutations (turquoise ◇ symbols) and 1 partially buried mutation (red *square*). The ten surface mutations rejected by the sorting procedure, of which two are included as green Δ symbols in Fig. 2, are indicated by their sequence position flanked by the mutated and mutant amino acids. The correlation coefficient between measured and computed $\Delta\Delta G$ on the 96 remaining surface mutations is equal to 0.87. With addition of the 24 partially buried and partially exposed mutations, it improves up to 0.89

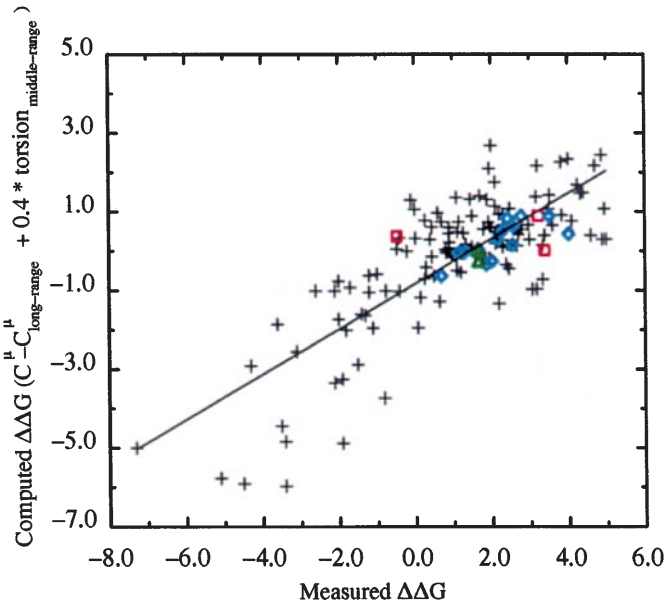

Fig. 2. $\Delta\Delta G$ computed with the sum of the $C^{\mu} - C^{\mu}_{\text{long-range}}$ potential and 0.4 times the $\text{torsion}_{\text{middle-range}}$ potential as a function of measured $\Delta\Delta G$ (in kcal/mol) for the set of 121 totally buried mutations (black + *symbols*), with in addition the subset of less-buried mutations that involve mainly non-local interactions along the chain, i.e. 2 surface mutations (green Δ symbols), 14 mutations of partially exposed mutations (turquoise ◇ symbols) and 3 partially buried mutations (red *squares*). The correlation coefficient between measured and computed $\Delta\Delta G$ on the 121 totally buried mutations is equal to 0.80. With addition of the 19 partially buried, partially exposed and fully exposed mutations, it drops to 0.79

a correlation coefficient between experimental and computed $\Delta\Delta G$s of 0.78 [7]. Thus we recover the well-known result that hydrophobicity is the dominating stabilising force in the protein core. This score of 0.78 is improved up to 0.80 by adding to this potential the $\text{torsion}_{\text{middle-range}}$ potential weighted by a factor of 0.4 (Fig. 2). This increase can be taken as statistically significant: when shuffling the $\Delta\Delta G$s computed with the torsion potential and adding them, with various weighting coefficients, to the $\Delta\Delta G$s computed with the distance potential, the score is increased by 0.02 or more in only 1 out of 1000 trials. Hence, though the hydrophobic interactions are the most important ones, the local interactions along the chain, responsible for secondary structure formation, are not negligible.

Though this score of 0.80 is not bad, it is significantly lower than the score of 0.87 obtained, with the sole torsion potential, for 96 out of the 106 surface residue substitutions. The $C^{\mu} - C^{\mu}$ potential therefore seems to measure the $\Delta\Delta G$s for core residues less well than the torsion potential does for surface residues. Several reasons can be invoked to explain this. One is related to packing modifications. Indeed, among the mutations of buried residues, some involve rather large changes in side chain size. They therefore modify the packing in the core: either the mutations create cavities or they induce strain. According to the flexibility in the environment of the mutations, the cavities are more or less easily filled and the strain relaxed. In principle, the distance potentials could take this effect into account, but they do not seem to be precise enough. To check this, we consider among the 121 core mutations the 23 mutations for which the radii of the mutated and mutant amino acids differ by at most 0.1 Å. On this subset, the correlation coefficient increases up to 0.87. It thus seems clear that one of the shortcomings of our procedure is that it does not account correctly for cavity formation and filling.

However, this is probably not the only shortcoming, as we shall now see. The $C^{\mu} - C^{\mu}_{\text{elec}}$ potential was not selected in the analysis described because it performs less well than the other $C^{\mu} - C^{\mu}$ potentials. On the 121 mutations of buried residues, it reaches a correlation coefficient of only 0.67 [8]. If we focus on the subset of 75 core mutations where both the mutated and mutant amino acids are hydrophobic, it does even worse, with a score of 0.22. In contrast, on the subset where the mutated or mutant (or both) amino acids are charged, the correlation coefficient improves up to 0.83. This potential is thus better suited than the other $C^{\mu} - C^{\mu}$ potentials to describe charge-charge interactions, but less suited to describe hydrophobic interactions. This indicates that there is not a single universal potential; the optimum potential depends not only on the solvent accessibility of the mutated residues and on the flexibility of the environment, but also on the amino acid types.

4.3 Mutations of partially buried residues (20% ≤ solvent accessibility < 40%)

For the 69 mutations of partially buried residues, neither torsion nor distance potentials taken alone yield good correlations between computed and measured $\Delta\Delta G$s [7]. To reach good scores, the potentials must be combined. The best correlation is obtained by adding the $C^{\mu} - C^{\mu}$ potential weighted by a factor of 0.7 to the $\text{torsion}_{\text{short-range}}$ potential (Fig. 3); however, the correlation coefficient so obtained is not very high (0.71). As in the case of the surface mutations, this is due to a few mutations that are far from the main group. Using our sorting procedure, we identify the outsiders: T26 → E in barnase, A41 → V in T4 lysozyme and D71 → A and V79 → G in chymotrypsin inhibitor. With these mutations excluded, the correlation coefficient is 0.82.

The reason why these mutations are far from the regression line seems to be that the interactions involved are not well described by the considered combination of potentials. Indeed, the first mutation fits well in the set of surface mutations (Fig. 1) and thus implies essentially local interactions along the chain, while the other three mutations fit well in the set of core mutations (Fig. 2) where hydrophobicity is preponderant.

4.4 Mutations of partially exposed residues (40% ≤ solvent accessibility < 50%)

None of the tested linear combinations of torsion and distance potentials leads to a good score for the 48 mutations of partially exposed residues [7]. They actually

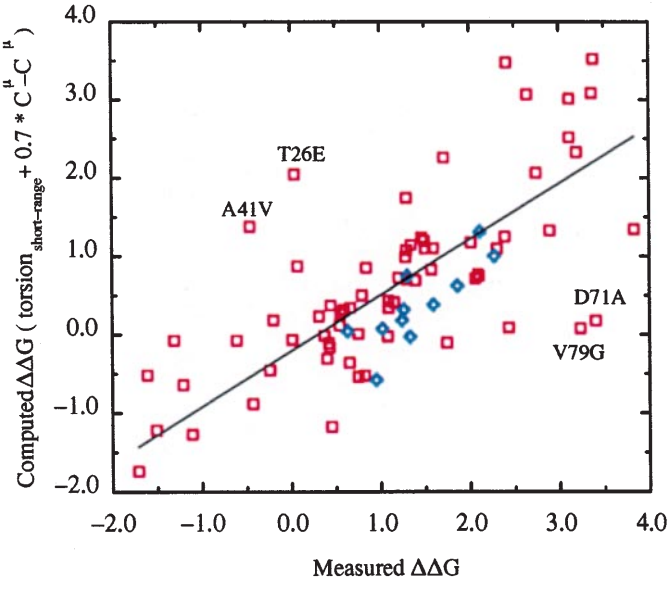

Fig. 3. $\Delta\Delta G$ computed with the sum of the $\text{torsion}_{\text{short-range}}$ potential and 0.7 times the $C^{\mu} - C^{\mu}$ potential as a function of measured $\Delta\Delta G$ (in kcal/mol) for the set of 69 partially buried mutations (red *squares*), with in addition the subset of mutations for which the local and non-local interactions along the chain have roughly the same weight, i.e. 11 mutations of partially exposed mutations (turquoise ◇ symbols). The four partially buried mutations rejected by the sorting procedure and included as red squares in Figs. 1 and 2 are indicated by their sequence position flanked by the mutated and mutant amino acids. The correlation coefficient between measured and computed $\Delta\Delta G$ on the 65 remaining partially buried mutations is equal to 0.82. With addition of the 11 partially exposed mutations, it drops to 0.80

seem to have no common characteristics. Indeed, these mutations can be divided into three subsets, which fit in the sets of mutations of surface, partially and totally buried residues, respectively (Figs. 1–3). This clearly shows that for these mutations, the solvent accessibility of the mutated residues is not a good measure for determining which potential must be used, and thus what the dominant interactions are.

5 Discussion

The first conclusion that emerges from our analysis is that when moving from the protein surface into the core, the non-local interactions along the chain, represented by distance potentials, gain importance, whereas the local interactions, represented by torsion potentials, lose importance, without actually disappearing. On the one hand, this confirms the role of hydrophobic interactions in the core; on the other hand, it emphasizes the dominance of local interactions at the surface and their non-negligible role in the core.

Our analysis also clearly shows the uselessness of trying to design a single universal database-derived potential able to reliably evaluate protein folding free energies. These energies can only be evaluated using linear combinations of different kinds of potentials, with coefficients depending, among other things, on the solvent accessibility of the mutated residues. This approach allows satisfactory estimation of stability changes caused by single-site mutations. Its accuracy should, however, improve by considering cavity formation or filling, by varying the combination of potentials as a function of the amino acid types, and by taking possible backbone structure modifications into account. The latter improvement should allow the approach to be extended to multiple mutations. This will be the objective of future work.

Acknowledgements. D. Gilis is Research Assistant at the Fonds pour la Formation a la Recherche dans l'Industrie et l'Agriculture (FRIA). Marianne Rooman is Senior Research Associate at the Belgian National Fund for Scientific Research (FNRS).

References

1. Fersht AR, Serrano L (1993) Curr Opin Struct Biol 3: 75
2. (a) Alber T, Daopin S, Wilson K, Wozniak JA, Cook SP, Matthews BW (1987) Nature 330: 41;
   (b) Matthews BW, Nicholson H, Becktel WJ (1987) Proc Natl Acad Sci USA 84: 6663;
   (c) Yutani K, Ogasahara K, Tsujita T, Sugino Y (1987) Proc Natl Acad Sci USA 84: 4441;
   (d) Kellis JT Jr, Nyberg K, Sali D, Fersht AR (1988) Nature 333: 784;
   (e) Matsumura M, Becktel WJ, Matthews BW (1988) Nature 334: 406;
   (f) Matoushek A, Kellis JT Jr, Serrano L, Fersht AR (1989) Nature 340: 122;
   (g) Serrano L, Fersht AR (1989) Nature 342: 296;
   (h) Daopin S, Baase WA, Matthews BW (1990) Proteins Struct. Funct. Genet. 7: 198;
   (i) O'Neil KT, DeGrado WF (1990) Science 250: 646;
   (j) Serrano L, Horovitz A, Avron B, Bycroft M, Fersht AR (1990) Biochemistry 29: 9343;
   (k) Daopin S, Alber T, Baase WA, Wozniak JA, Matthews BW (1991) J Mol Biol 221: 647;
   (l) Sali D, Bycroft M, Fersht AR (1991) J Mol Biol 220: 779;
   (m) Eriksson AE, Baase WA, Zhang X-J, Heinz DW, Blaber M, Baldwin EP, Matthews BW (1992) Science 255: 178;
   (n) Horovitz A, Matthews JM, Fersht AR (1992) J Mol Biol 227: 560;
   (o) Hu C-Q, Kitamura S, Tanaka A, Sturtevant JM (1992) Biochemistry 31: 1643;
   (p) Serrano L, Kellis JT Jr, Cann P, Matouschek A, Fersht AR (1992) J Mol Biol 224: 783;
   (q) Serrano L, Sancho J, Hirshberg M, Fersht AR (1992) J Mol Biol 227: 544;
   (r) Zhang X-J, Baase WA, Matthews BW (1992) Protein Sci 1: 761;
   (s) Blaber M, Zang X, Matthews BW (1993) Science 260: 1637;
   (t) Jackson SE, Moracci M, el Masry N, Johnson CM, Fersht AR (1993) Biochemistry 32: 11259;
   (u) Matthews SJ, Jandu SK, Leatherbarrow RJ (1993) Biochemistry 32: 657;
   (v) Jackson SE, Fersht AR (1994) Biochemistry 33: 13880;
   (w) Itzhaki LS, Otzen DE, Fersht AR (1995) J Mol Biol 254: 260;
   (x) Otzen DE, Fersht AR (1995) Biochemistry 34: 5718;
   (y) Shih P, Holland DB, Kirsch JF (1995) Protein Sci 4: 2050;
   (z) Shih P, Kirsch JF (1995) Protein Sci 4: 2063
3. (a) Shoichet BK, Baase WA, Kuroki R, Matthews BW (1995) Proc Natl Acad Sci USA 92: 452;
   (b) Takano K, Ogasahara K, Kaneda H, Yamagata Y, Fujii S, Kanaya E, Kikuchi M, Oobatake M, Yutani K (1995) J Mol Biol 254: 62;
   (c) Zhang X-J, Baase WA, Shoichet BK, Wilson KP, Matthews BW (1995) Protein Eng 8: 1017;
   (d) Kay MS, Baldwin RL (1996) Nat Struct Biol 3: 439
4. (a) Basch PA, Singh UC, Langridge R, Kollman PA (1987) Science 236: 564;
   (b) Tidor B, Karplus M (1991) Biochemistry 30: 3217.
5. (a) Lee C, Levitt M (1991) Nature 352: 448;
   (b) Koehl P, Delarue M (1994) Proteins Struct. Funct. Genet. 20: 264;
   (c) Lee C (1994) J Mol Biol 236: 918;
   (d) Muñoz V, Serrano L (1994) Proteins Struct. Funct. Genet. 20: 301;
   (e) Miyazawa S, Jernigan RL (1994) Protein Eng 7: 1209;
   (f) Sippl MJ (1995) Curr Op in Struct Biol 5: 229;
   (g) Ota M, Shigenori K, Nishikawa K (1995) J Mol Biol 248: 733;
   (h) Topham CM, Srinivasan N, Blundell TL (1997) Protein Eng 10: 7;
   (i) Damborsky J (1998) Protein Eng 11: 21
6. Gilis D, Rooman M (1996) J Mol Biol 257: 1112
7. Gilis D, Rooman M (1997) J Mol Biol 272: 276
8. Rooman M, Gilis D (1998) Eur J Biochem 254: 135
9. Alard P (1991) PhD thesis. Université Libre de Bruxelles
10. Wintjens RT, Rooman MJ, Wodak SJ (1996) J Mol Biol 255: 235
11. (a) Rooman MJ, Kocher J-PA, Wodak SJ (1991) J Mol Biol 221: 961;
    (b) Rooman MJ, Kocher J-PA, Wodak SJ (1992) Biochemistry 31: 10226
12. Kocher J-PA, Rooman MJ, Wodak, SJ (1994) J Mol Biol 235: 1598