Protein Engineering vol.10 no.1 pp.7–21, 1997

Prediction of the stability of protein mutants based on structural environment-dependent amino acid substitution and propensity tables

Christopher M.Topham[1], N.Srinivasan[3] and Tom L.Blundell[2,3]

ICRF Unit of Structural Molecular Biology, Department of Crystallography, Birkbeck College, University of London, Malet Street, London WC1E 7HX, UK

[1]Present address: Cancer Drug Discovery, Department of Chemistry, University College Dublin, Belfield, Dublin 4, Ireland
[3]Present address: Department of Biochemistry, Tennis Court Road, Cambridge CB2 1QW, UK

[2]To whom correspondence should be addressed

An approach to the prediction of mutant stability is described using knowledge of amino acid replacements that are tolerated within the families of homologous proteins of known 3-D structure. Amino acid variations in families of homologous proteins are converted to propensity and substitution tables; these provide quantitative information about the existence of an amino acid in a structural environment and the probability of replacement by any other amino acid. The tables are used to calculate a 'stability difference score', analogous to the difference in free energy between a mutant and the wild type. The method has been developed and tested using the high-resolution structures for T4 lysozyme and 159 site-specific mutants. We show that differences in stability scores are correlated with experimentally observed free energy differences and differences in melting temperature. Blind tests, using only structural information derived from the parent wild-type crystal structures, on a combined set of 83 staphylococcal nuclease and 68 barnase mutants showed a correlation of 0.80 in the predicted stability changes with experimental thermodynamic data. Approximately 86% of the predictions were correctly classified as destabilizing or stabilizing.

Keywords: environment-dependent amino acid substitution/point mutations/prediction/propensity tables/protein stability

Introduction

Interest in the prediction of protein thermal stability has been stimulated by site-directed mutagenesis which allows the stability of a protein to be investigated by the substitution of individual amino acids (Alber, 1989; Fersht and Winter, 1992; Matthews, 1993). Energy differences between such mutants and the wild type have been estimated using a variety of computer simulation methods. A Monte Carlo simulated annealing protocol, devised by Lee and Levitt (1991) and extended by Lee (1994), has been used to obtain a correlation between theoretical stabilization energies and thermal stability for a series of nine core mutants of λ repressor. However, the use of a simplified force field in conjunction with an elaborate search algorithm to identify low energy mutant side-chain packing configurations has been criticized by van Gunsteren and Mark (1992), who obtained comparable results based on relatively straightforward structural considerations. 'Alchemical' molecular dynamics simulations (Goa *et al.*, 1989; Straatsma and McCammon, 1992), employing more accurate potential functions, have also been used to determine the effect of mutations on protein stability (Bash *et al.*, 1987; Wong and McCammon, 1987; Dang *et al.*, 1989; Prévost *et al.*, 1991; Tidor and Karplus, 1991; Sneddon and Tobias, 1992). However, the application of free energy perturbation and thermodynamic integration methods is very computer intensive. The reliability of such calculations may also be compromised by finite simulation times and the choice of operational protocol (Shi *et al.*, 1993).

A computationally more economic alternative to simulation seeks to exploit the increasingly large body of knowledge contained within protein databases concerning the forces that stabilize proteins. In common with simulation techniques, knowledge-based predictions necessarily involve either making explicit assumptions about the physical nature of the unfolded state or implicitly assuming that energetic contributions to estimates of relative thermal stabilities originating from the unfolded state cancel one another out. Bordo and Argos (1990) suggested residue substitutions for structural stability by analysing the characteristics of mutated residues in the core of structurally aligned globins. Subsequently this work was extended to analyse natural mutations in interacting residue pairs that share a conserved environment across the members of a family (Bordo and Argos, 1991). The exchange matrices derived provided guidelines for 'safe' amino acid substitutions least likely to disturb the tertiary structure. Miyazawa and Jernigan (1994) have described an empirical method based on effective inter-residue contact energies and an assumption about the compactness of the unfolded state. Jones (1994) has used pairwise knowledge-based potentials and a genetic algorithm to identify mutations that are compatible with a given fold. Recently Ota *et al.* (1995) used a pseudo-energy potential combining side-chain packing, hydration, hydrogen bonding fitness and local conformation terms to obtain a correlation of 0.51 when stability scores are plotted against changes in the melting temperature of 81 *Escherichia coli* ribonuclease H1 mutants. The average energy of an ensemble of residue environments, derived from a protein structure database, was taken to represent the energy of the denatured state for a given residue type. Based on the theoretical arguments presented by Rooman and Wodak (1995), Gilis and Rooman (1996) employed a similar representation of the unfolded state and used backbone torsion angle propensities and distance-dependent residue–residue potentials to predict free energy changes in surface mutants. Sippl (1995) has also reported the application of a knowledge-based mean force potential, comprising pair interactions among all backbone atoms plus a term for protein–solvent interactions, to multiple amino acid replacements at position 32 in barnase and position 44 in T4 lysozyme.

In this study we introduce a method for the rapid prediction of relative mutant protein stabilities based on statistical analyses

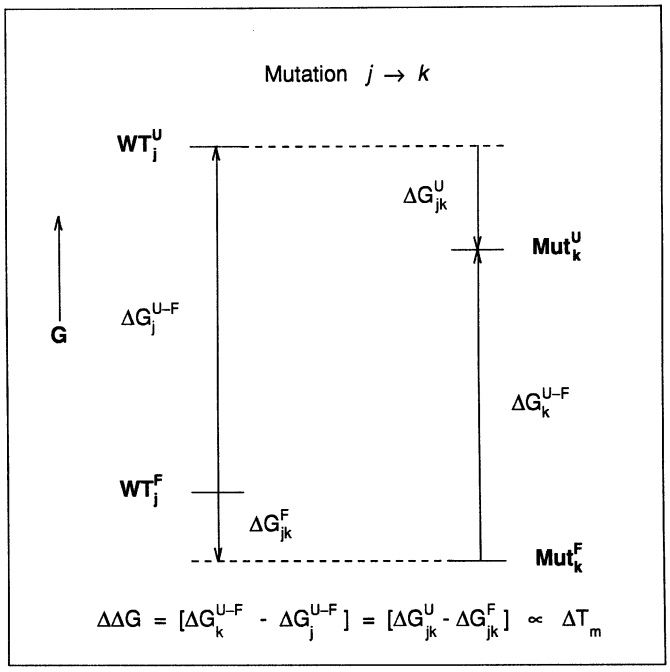

Fig. 1. Folding–unfolding free energy diagram for site-directed mutagenesis.

of amino acid replacements that are tolerated within the families of homologous proteins of known 3-D structure. The problem of energetic contributions from the unfolded state is addressed by considering subsets of data for exposed, non-hydrogen bonded residues in regions outside the secondary structure. The method is tested by comparing a 'stability difference score', analogous to the free energy difference between a mutant and the wild-type protein, with experimental thermal stability data for a wide range of engineered bacteriophage T4 lysozyme mutants for which X-ray structures are available. The high level of correlation with experimentally observed data was reproduced in tests with staphylococcal nuclease and barnase point-site mutants using structural information derived only from their parent wild-type crystal structures.

Materials and methods

Background

Protein stability is dependent on the difference in free energy (ΔG^{U-F}) between the folded (F) and unfolded (U) states ($\Delta G^{U-F} = G^U - G^F$). The mutation of residue type j in the wild-type protein to residue type k may be coupled to the reversible folding–unfolding process by means of the familiar thermodynamic cycle in Figure 1. $\Delta\Delta G$, the difference in free energy of unfolding of the wild type and mutant, is related to ΔG_{jk}^U and ΔG_{jk}^F, the respective free energy changes associated with the transformation of $j \rightarrow k$ in the unfolded and folded states through the thermodynamic identity

$$\Delta\Delta G = (\Delta G_k^{U-F} - \Delta G_j^{U-F}) = \Delta G_{jk}^U - \Delta G_{jk}^F \qquad (1)$$

$\Delta\Delta G$ can also be approximately related to small differences in melting temperatures, ΔT_m (Becktel and Schellman, 1987)

$$\Delta\Delta G \approx \Delta S_{T_{m(j)}}^* \cdot \Delta T_m, \qquad (2)$$

where $\Delta T_m = T_{m(k)} - T_{m(j)}$ and $\Delta S_{T_{m(j)}}^*$ is the change in entropy at the melting temperature of wild type. Theoretical predictions of thermal stabilities of mutant proteins may therefore be tested by comparison of calculated values of $\Delta G_{jk}^U - \Delta G_{jk}^F$ with experimentally determined values of $\Delta\Delta G$ and ΔT_m.

Calculation of stability changes in mutant proteins

Our knowledge-based approach to the calculation of changes in the thermal stability associated with the replacement of residue type j with residue type k is based on conformationally constrained environment-dependent amino acid substitution and propensity tables described previously (Topham *et al.*, 1993, 1994). These tables were constructed from a compilation of observed frequencies of occurrence and substitution in an alignment database (Overington *et al.*, 1990, 1992; Šali and Overington, 1994). The database (version 3B) comprised 506 protein–protein domains in 113 families, following the removal of the photosynthetic reaction centre and porin membrane protein families. Several structural descriptors at each residue position were optimally aligned (Šali and Blundell, 1990) for proteins in each family. The combination of main-chain conformations (nine classes), solvent accessibilities (three classes) and side-chain hydrogen bonding with other side chains, as well as the amides and carbonyls of the main chain (two or eight classes), accounts for a total possible 216 environment types when eight (2^3) combinations of hydrogen bonding are considered, or 54 types when side chains are simply taken to be involved or not in hydrogen bond formation. Distinction is made between half-cystine and cysteine residues, resulting in 21 residue types.

A total of 345 180 substitutions were observed in which the main-chain conformational class is conserved. These are classified into 54 (9×3×2) structural environments (ε). The observed substitution frequencies are corrected for relative amino acid mutability and normalized frequency of residue type occurrence in the global data set, summed over all 54 environmental classes, and converted to conditional probabilities (Topham *et al.*, 1993). To reduce bias on the diagonal of the tables, observed frequencies were also weighted according to sequence dissimilarity of the proteins in the pairwise comparisons within a family. An entropy-driven (partial) smoothing procedure was employed to minimize the problem of sparse data, as described by Topham *et al.* (1993).

The propensity table, derived by Topham *et al.* (1994), may be viewed as an extended version of the (18) environment×(20) residue type table described by Bowie *et al.* (1991). In the present work, a 216 (9×3×8) environment×21 residue propensity table was created from 64 616 observations of unique residue environment combinations at positions in the structural alignments. Conditional probabilities of finding a particular residue type in a given environment were calculated by correcting for the normalized frequency of residue type occurrence in the global data set. Smoothing of the environment-dependent probabilities of residue occurrence was also carried out.

Two somewhat similar approaches have been described in the literature for the derivation of knowledge-based potentials from observed frequencies of occurrence of amino acid sequence and protein structural features in databases (Rooman and Wodak, 1995). Statistical mechanics treatments in which potential terms are expressed as $-kT \cdot \ln$ (probability ratios), where k is Boltzmann's constant and T a temperature, lead to expressions for the derived potential that comprise the free energy of a given conformer plus a term representing the partition function of the system (Sippl, 1990, 1993, 1995; Wodak and Rooman, 1993; Rooman and Wodak, 1995). Whilst

in principle this approach permits the comparison of different conformations of a given sequence, energies of different sequences cannot be compared directly without making an additional assumption effectively equating the partition functions of all sequences. For this reason, when making predictions of mutant protein relative stability we prefer to work in the alternative context of a purely statistical model of dimensionless 'stability scores' (s), expressed as ln (probability ratios). Stability score differences (Δs) are used to represent the component processes of the folding–unfolding cycle of Figure 1. Thus, by analogy with Equation 1, we can write

$$\Delta\Delta s = \Delta s^{U}_{jk} - \Delta s^{F}_{jk}, \tag{3}$$

where Δs^{F}_{jk} and Δs^{U}_{jk} are differences in stability scores associated with the non-physical interconversion of residue types j and k in the folded and unfolded states, respectively. As $\Delta\Delta s$ and $\Delta\Delta G$ do not share the same dimensions, their magnitudes cannot be equated. However, their signs and rank order should agree and, in this sense, provide a direct index of protein stability. By establishing a statistical correlation between $\Delta\Delta s$ and $\Delta\Delta G$ (or) ΔT_m, it is possible to obtain estimates of $\Delta\Delta G$ ($\Delta\hat{\Delta}G$) and ΔT_m ($\Delta\hat{T}_m$) from calculated values of $\Delta\Delta s$. These may then of course be compared with experimentally determined values.

Two different methods are used to calculate the Δs^{U}_{jk} and Δs^{F}_{jk} terms in Equation 3. In Method I, stability difference scores are derived from environmental substitution data, and in Method II, from propensities. The difference in information content of environmental substitution and propensity tables is discussed by Topham *et al.* (1993). In considering the transformation of residue j to residue k in the folded state, both methods require a knowledge of the physical environments (ε) of the respective residue types in the wild-type protein and in the mutant. This is done to ensure that estimates of relative stability are independent of whether the wild type or the mutant is chosen as the starting configuration. The elements contained within a given substitution table, corresponding to one of 54 structural environments, refer to the probability of a residue type R_j existing in that particular environment being replaced by a residue type r_k in any environment. Thus, as the substitution tables take into account only the environment of one of the two residues, it is necessary to consider two probabilities: namely that of the replacement of residue type R_j in the wild-type environment (ε_{wt}) by residue type r_k in an undefined environment [$P(r_k/R_j,\varepsilon_{wt})$], and the probability of the replacement of residue type R_k in the mutant environment (ε_{mut}) by residue type r_j in an undefined environment [$P(r_j/R_k,\varepsilon_{mut})$]. In calculating the ratio of these probabilities, it is not necessary to correct for the respective global environment-independent substitution probabilities [$P(r_k/R_j)$, $P(r_j/R_k)$] because they are equal ($1/21 = 0.046$) and therefore cancel out. However, it is apparent that the column vectors of the substitution matrices contain redundant information about the probability of replacement of residue type R_j by residue types other than r_k. Thus, in order to avoid bias when combining probabilities derived from different column vectors, it is necessary to re-normalize by the introduction of a reference state. A suitable choice in the case of residue type R_j in the wild-type environment is the probability of it being conserved, or reverting to self (r_j), in any environment [$P(r_j/R_j,\varepsilon_{wt})$]. The reference state in the case of residue type R_k in the mutant environment is the analogous conditional probability $P(r_k/R_k,\varepsilon_{mut})$. Combination of the four probability terms and summing over point mutations at the ith residue position yields

$$\Delta s^{F}_{jk} = \sum_i -\ln \left\{ \frac{P(r_{ki}/R_{ji},\varepsilon_{wt})}{P(r_{ji}/R_{ji},\varepsilon_{wt})] } \cdot \frac{P(r_{ki}/R_{ki},\varepsilon_{mut})}{P(r_{ji}/R_{ki},\varepsilon_{mut})} \right\} \tag{4}$$

The difference in stability scores between the two unfolded ground states Δs^{F}_{jk} can be derived in an analogous way. The question arises as to what are the environments of residues j and k in the unfolded state. Privalov *et al.* (1989) have shown that the heat capacities, intrinsic viscosities and CD properties of a number of unfolded proteins under most conditions have very little regular secondary structure, although there is evidence that residual structure can be present in unfolded proteins (Dill and Shortle, 1991; Evans *et al.*, 1991; Pace *et al.*, 1992). We have used a conformationally constrained substitution table for non-hydrogen bonded, surface-exposed amino acid residues falling outside regions of regular secondary structure. The score difference, $\Delta\Delta s$, can then be obtained according to Equation 3.

The second method for calculating Δs^{U}_{jk} and Δs^{F}_{jk} from amino acid environmental propensities is conceptually simpler. In considering the folded state, the propensity table is used to define the relative probabilities of finding residue types j [$P(r_j/\varepsilon_{wt})$] and k [$P(r_k/\varepsilon_{mut})$] in their respective environments in the wild-type and mutant structures

$$\Delta s^{F}_{jk} = \sum_i -\ln \left\{ \frac{P(r_{ki}/\varepsilon_{mut})}{P(r_{ji}/\varepsilon_{wt})} \right\} \tag{5}$$

where i refers to the residue position of a point-site mutation. The relative probability of finding residue types j and k in any environment [$P(r_j)/P(r_k)$] is simply unity, and so correction of Equation 5 is not necessary. Δs^{U}_{jk} was calculated using the row vector for non-hydrogen bonded, surface-exposed amino acid residues falling outside regions of regular secondary structure. As before, $\Delta\Delta s$ is calculated using Equation 3.

It is evident that some laboratory engineered proteins, such as cavity mutants, are without natural counterparts. This is because the families of homologous proteins which were compared have accepted multiple mutations, and evolutionary pressure has generally aided so as to fill the vacated space. The effects of single substitutions are not often observed over the timescale of evolution. To compensate for this, we introduce a disruption term. Equation 3 becomes:

$$\Delta\Delta s = \Delta s^{U}_{jk} - \Delta s^{F}_{jk} - \Delta s^{Disrupt}_{jk}. \tag{6}$$

The disruption term was applied only to buried mutated residues. It is defined as a logarithmic function of the absolute value of the net change over point-site mutation positions (i) in the side-chain surface-accessible area (A_X) in an extended peptide Gly–X–Gly, relative to that for glycine.

$$\Delta s^{Disrupt}_{jk} = \ln \left\{ 1 + \left| \sum_i \frac{(A_{ji} - A_{ki})}{A_{Gly}} \right| \right\} \tag{7}$$

Testing and application of the predictive model

In developing the prediction method, we have used crystallographic structures of bacteriophage T4 lysozyme mutants solved by Matthews *et al.* (1987) and available in the January 1995 release of the Brookhaven Protein Data Bank (PDB; Bernstein *et al.*, 1977). The four-letter PDB codes of the mutant structures are 1L02–1L09, 1L11–1L15 (Alber *et al.*,

1987), 1L25–1L32 (Alber *et al.*, 1988), 1TLA (Anderson *et al.*, 1993), 140L–147L (Baldwin *et al.*, 1993), 1L34 (Becktel and Schellman, 1987), 1LYE–1LYJ (Bell *et al.*, 1992), 118L–128L (Blaber *et al.*, 1993), 107L–115L, 216L, 137L, 1DYA–1DYG (Blaber *et al.*, 1994), 155L–166L (Blaber *et al.*, 1995), 1L58 (Chen *et al.*, 1992), 1L33 (Dao-pin *et al.*, 1990), 1L37–1L41 (Dao-pin *et al.*, 1991a), 1L48–1L53 (Dao-pin *et al.*, 1991b), 1L42–1L47 (Dao-pin *et al.*, 1991c), 1L54 (Dao-pin *et al.*, 1991d), 1L96, 1L97 (Dixon *et al.*, 1992), 1L63, 1L69 (Eriksson *et al.*, 1992a), 1L83, 1L84 (Eriksson *et al.*, 1992b), 1L85–1L95 (Eriksson *et al.*, 1993), 150L (Faber and Matthews, 1990), 1L16 (Gray and Matthews, 1987), 1L10 (Grütter *et al.*, 1987), 1L64–1L68 (Heinz *et al.*, 1992), 1L77, 2L78, 1L79–1L82 (Hurley *et al.*, 1992), 1L17, 1L18, 149L (Matsumura *et al.*, 1988), 1L35 (Matsumura *et al.*, 1989), 1L23, 1L24 (Matthews *et al.*, 1987), 1L19, 1L20 (Nicholson *et al.*, 1988), 1L21, 1L22 (Nicholson *et al.*, 1989), 1L55, 1L57, 1L59, 1L61, 1L62 (Nicholson *et al.*, 1991), 1L56, 1L60 (Nicholson *et al.*, 1992), 1L98, 1L99, 1L00 (Pjura *et al.*, 1993a), 129L–131L (Pjura *et al.*, 1993b), 1L76 (Sauer *et al.*, 1992), 1L36, 1L70, 1L71 (Zhang *et al.*, 1991) and 1L72–1L75 (Zhang *et al.*, 1992). Residue environments were determined at each position in the fold and calculations of $\Delta\Delta s$ were made using Equation 3 and its extended form, Equation 6. Experimental values of $\Delta\Delta G$ and/or ΔT_m for 159 of the coordinate sets deposited could be found in the literature and have been collated at our anonymous ftp site (ftp.cryst.bbk.ac.uk). In many cases experimental estimates were available at strongly acidic and weakly acidic–neutral pH values. Both values were used in the regression analyses of the correlation between $\Delta\Delta s$ and the experimental data. The pseudo-wild-type (C54T,C97A) coordinate set (PDB code 1L63) was used in place of the wild-type protein whenever the mutant of interest also contained these mutations, with the exception of one case (PDB code 1L35) involving the mutation L164C. Atomic coordinates for Leu164 are missing from the 1L63 data set and the 3LZM set was used instead, but excluding contributions from positions 54 and 97. In all other cases the coordinate set chosen was the one used as the starting model in the refinement of the mutant structures (3LZM or 2LZM as appropriate).

To apply the method, physical environments of mutant residues could be either extrapolated from the wild type or, alternatively, determined directly from the coordinates of a model. A basic assumption of the extrapolation procedure is that residue replacement is not accompanied by a change in conformational class. Care was taken to ensure that incompatible mutant side-chain hydrogen bonding combinations were not introduced as a result of environment extrapolation. Thus engagement in side-chain hydrogen bond formation by amino acids that lack hydrogen bonding functionalities was disallowed. Extant wild-type hydrogen bonding combinations involving hydrogen bond formation to main-chain (–NH) functions were also removed when residues were replaced by Arg, Lys or Trp, as were apparent mutant hydrogen bonding combinations involving Met and main-chain (>C=O) functions. Predictions of mutant thermal stabilities calculated on the basis of wild-type environment extrapolations were compared with predictions based on the T4 lysozyme crystal structure environments using the same wild-type coordinate data sets.

Additional tests of the application of the prediction method were carried out on 68 barnase mutants, for which experimental $\Delta\Delta G$ values are tabulated in compendia by Serrano *et al.* (1992) and Matoushek *et al.* (1994), and 83 staphylococcal nuclease mutants (Shortle *et al.*, 1990). The barnase mutants were: I4V; I4A; N5A; T6G; T6A; D8A; D8A,D12A; D8A,V10A; D8A,D12A,R110A; V10T; V10A; D12A; D12A,R110A; Y13A; L14A; Q15I; T16S; T16R; Y17A; Y13A,Y17A; T16A,Y17A; T16S,Y17A; H18Q; N23A; Y24F; I25V; I25A; T26G; T26A; K27G; E29G; Q31S; Q31A; L33Q; V36A; V36T; N41D; V45A; V45T; I51V; I51A; D54A; D54N; I55V; I55A; V55T; N58A; N58D; K62R; I76V; I76A; N77A; Y78F; N84A; I88V; I88A; L89V; L89T; S91A; S92A; I96V; I96A; T99V; Y103F; T105V; I109V; I109A; R110A. The staphylococcal nuclease data set comprised five isoleucine to valine mutations (I15V; I18V; I72V; I92V; I139V), plus mutations of the following residues to alanine and glycine: L7, L14, I15, I18, V23, L25, M26, Y27, M32, F34, L36, L37, L38, V39, V51, Y54, F61, M65, V66, I72, V74, F76, Y85, L89, Y91, I92, Y93, M98, V99, L103, V104, L108, V111, Y113, V114, Y115, L125, L137 and I139. Collated values of $\Delta\Delta G$ for all the barnase and staphylococcal nuclease mutants are provided as supplementary material at our anonymous ftp site.

With the exception of a few barnase mutant crystal structures, coordinate data sets for these mutants are not available. Thermal stability predictions were made using the environment extrapolation procedure or were based on modelled structures. The barnase PDB 1BGS (A chain) (Guillet *et al.*, 1993) coordinate set was used as the wild-type template following removal of the atomic coordinates of the bound C40A,C82A mutant barstar molecule. The 1STN coordinate set (Hynes and Fox, 1991) was used as the staphylococcal nuclease wild-type. Models were built of the barnase and staphylococcal nuclease mutants using the side-chain mutation facility within the SYBYL 6.1 molecular graphics package (Tripos Inc.). The algorithm conserves as far as possible the side-chain torsion angles of the wild type. All main-chain and non-mutated side-chain atoms were held fixed. Minor side-chain geometry fixes were necessary for some mutants. These mainly involved rotations around the C^{α}–C^{β} bonds of Val and Thr residues constructed from wild-type Ile and Val residues, necessitated by conflicts in standard stereochemical χ^1 definitions. Adjustments were also made to the arginine side-chain conformation of the barnase K62R mutant to avoid a non-bonded contact between $N^{\eta 2}$ and the carbonyl oxygen of Gln104.

Statistical analysis

A linear regression analysis of the dependence of $\Delta\Delta s$, or estimates of $\Delta\Delta G$ ($\Delta\hat{\Delta}G$) obtained by calibration, on experimental data were performed using a robust two-stage procedure described previously by Topham *et al.* (1993). The procedure is designed to minimize the influence of outliers without their explicit removal. The first stage involves the calculation of estimates of the slope and ordinate intercept using Theil's non-parametric regression method as described by Sprent (1989). These are then used to seed the iterative bi-weight least-squares regression algorithm of Mosteller and Tukey (1977) in the second stage. An updated set of weights $[w_i^{(k+1)}]$ is defined as a function of scaled deviations $[u_i^{(k)}]$ obtained from the kth fit. The scaled deviations are defined as

$$u_i^{(k)} = \frac{e_i^{(k)}}{c \cdot S^{(k)}} \qquad (8)$$

where $c = 6$ and $S^{(k)}$ is the median of the absolute values of

the residuals, $|e_i^{(k)}|$, produced by the k^{th} fit. The weights in the next iteration are calculated as:

$$w_i^{(k+1)} = \begin{cases} 1 - u_i^{(k)2})^2 & \text{if } |u_i^{(k)}| \leqslant 1 \\ 0 & \text{if } |u_i^{(k)}| > 1 \end{cases}$$

Thus, when $|u_i^{(k)}|$ is moderate or small, $w_i^{(k+1)} \approx w_i^{(k)}$ and the weights remain unchanged, but if $|u_i^{(k)}|$ is large, $w_i^{(k+1)} << w_i^{(k)}$, and in the limit, zero. For distributions near the normal, $6S$ will be on average about 4σ. Convergence was reached in all cases within 10 iterations.

Residuals from the regression fits were examined using the *t* test criterion. Values of *t* were calculated using the formulation of Snedecor and Cochran (1980) which avoids the necessity of recomputing the regression line with each point omitted in turn. For each plot, the fraction of points with $t > 1.96$ was recorded as a function of the wild-type residue environments at mutation sites. Data points with *t* values >1.96 are associated with a 5% probability (for an infinite number of degrees of freedom) of belonging to the population regression line. Six residue environment classes were considered for purposes of residual analysis: involvement (or otherwise) in side-chain hydrogen bonding; buried (<7% relative side-chain accessibility) or partially buried surface (≥7% relative side-chain accessibility); within secondary structural element (helix/sheet) or not (coil). Data size limitations did not permit the consideration of environment combinations. In cases of multiple point mutation sites, each residue was taken to have contributed equally to a single estimate of relative stability.

Mutations were also categorized as being stabilizing or destabilizing, and the percentage agreement between the computed and experimental results calculated as a function of the same environments used in the analysis of residuals. The few mutants reported in the literature as being exactly equally stable as the parent wild type were assigned to the more stable category.

Results and discussion

Algorithm development and performance using T4 lysozyme mutant test set

The results obtained using both prediction methods for a representative set of T4 lysozyme mutant crystal structures are presented in Table I. Tabulated results for the entire data set can be obtained from the anonymous ftp site. Plots of $\Delta\Delta s^{I}$ and $\Delta\Delta s^{II}$ versus experimentally determined values of $\Delta\Delta G$ (Figures 2 and 3) demonstrate significant correlation over a broad range of environments. A total of 217 data points were included in the regression and residual analyses. Broken down according to environment, 31.5% of the mutations are at buried sites, 35.0% are involved in side-chain hydrogen bond formation, and 23.8% of residue replacements occur outside of secondary structural regions. The correlation coefficient (*r*) increases from 0.56 to 0.65 for Method I, based on environmental amino acid substitution data, when the disruption term is included at buried residue sites. This is accompanied by a drop in the percentage of outlying data points with values >1.96 from 6.6 to 4.6% (Table II). A smaller improvement in the correlation coefficient from 0.72 to 0.77 is seen when the predictions are based on environmental propensities (Method II) with no change in the percentage of data points (3.7%) with values >1.96 (Table II). When the disruption term is applied, 65.4% of the predictions are correctly assigned as stabilizing or destabilizing using Method I and 73.3% using Method II (Table II). Examination of the statistical data in Table II as a function of the wild-type residue environment reveals that mutations at buried residue positions are the best predicted (75.9 and 83.7% respective agreement for Methods I and II); the least well predicted are mutations of residues involved in hydrogen bonding (53.2%, Method I) and mutations in coil regions (63.2%, Method II). There were no examples of outlying predictions, associated with *t* values >1.96, that led to incorrect stability assignments by both methods.

A significant proportion of the T4 lysozyme mutations lead to an increase in thermal stability (29.5%). Of these, 79.7% are successfully predicted by Method I as having positive $\Delta\Delta s$ values, and 82.8% by Method II. One way in which protein stability may be enhanced is by the introduction of an acidic group close to the N-terminus of an α-helix. Examples of such stabilizing mutations include S38D and N144D (Nicholson *et al.*, 1988), T109D and N116D (Nicholson *et al.*, 1991), and T115E (Dao-pin *et al.*, 1991a). Each substitution increases stability at neutral pH where the substituting residue is charged, but not at low pH where it is protonated and uncharged. Positive values are obtained for all five mutants irrespective of whether the predictions are based on environmental amino acid substitutions or propensities (Table I). Structural studies indicate that in the case of N144D, shown in Figure 4, and the geometrically similar T109D and N116D N-cap +1 α-helix substitutions, the mutant side chains do not make hydrogen bonds to the end of the helix. This is consistent with stabilization by a generalized electrostatic interaction with the helix dipole, the entropic cost of localizing the partial positive charge at the N-terminus of the helix being provided during protein folding (Matthews, 1993).

The use of the disruption term at buried sites is primarily intended to compensate for the destabilizing effect of cavity creation in engineered mutants without natural counterparts. Eriksson *et al.* (1992a) found an approximately linear relationship between loss in protein stability and the size of cavity created in cavity-containing T4 lysozyme mutants with truncated hydrophobic side chains. The decrease in stability associated with Leu → Ala replacements, for example, was interpreted in terms of a constant free energy term of 1.9 kcal/mol, equivalent to the difference in solvent transfer-free energy from water to octanol for leucine and alanine, plus a variable component that depends on the size of the cavity created by the substitution. The cavity size increases in the Leu → Ala mutants investigated by Eriksson *et al.* (1992a) range from 24.3 (L46A) to 122.5 Å^3 (L99A), contributing 0.58–2.94 kcal/mol towards the free energy difference, based on a cavity-dependent energy cost term of 0.024 kcal/mol/Å^3. This corresponds to a contribution of 59% to the overall energy change in the case of L99A at pH 3.01, and 22% for L46A. The formulation of the disruption term used in the predictive model (Equation 7) is based on the simplifying assumption that cavity size change would depend on the net change in side-chain solvent-accessible area buried by the relevant side chains involved. Thus $\Delta s_{jk}^{Disrupt}$ is a fixed penalty term, dependent only on the identities of the amino acid types *j* and *k*. The L133A mutant with an intermediate cavity size change of 77.8 Å^3, responsible for 52% of the overall change of 3.6 kcal/mol, serves as a useful reference to assess the magnitude of the disruption term contribution. Both methods correctly predict L133A to be less stable than the wild type (Table I). The

Table I. Selected T4 lysozyme mutant structures illustrating the thermal stability prediction algorithm

WT PDB code	Mutant PDB code	Mutation	$\Delta s_{jk}^{\text{Disrupt.}}$	Method I Δs_{jk}^{U}	Method I Δ_{jk}^{F}	Method I $\Delta\Delta s^{I}$	Method II Δs_{jk}^{U}	Method II Δ_{jk}^{F}	Method II $\Delta\Delta s^{II}$	Av. est. $\Delta\hat{\Delta}G$ [*kcal*	Expt. $\Delta\Delta G$ mol^{-1}]	Av. est. $\Delta\hat{T}_m$ [Ω	Expt. ΔT_m	pH	Comment	Reference
3LZM	1L97A	I3P	0.0	–4.5	–2.9	–1.5	–1.4	0.5	–1.9	–4.0	–2.8	–13.4	–7.3	3.01	Proline replacement	Dixon *et al.* (1992)
3LZM	1L17	I3V	0.0	–0.2	0.6	–0.8	–0.2	0.2	–0.4	–1.8	–0.6	–6.0	–2.1	2.0	Hydrophobic replacement	Matsumura *et al.* (1988)
											–0.4		–1.2	6.5		
3LZM	1L35′	I9C,L164C,WT*	0.0	–9.3	–9.8	0.4	1.2	1.0	0.2	–0.2	1.5	–0.5	6.4	2.0	Stabilizing S-S bridge	Matsumara *et al.* (1989)
													–1.5	2.8		
3LZM	1L42	K16E	0.0	0.5	0.5	0.0	0.3	0.4	–0.1	–0.8	–0.5	–2.8	1.1	5.3	Cumulative charge–charge	Dao-pin *et al.* (1991c)
											0.5					
3LZM	1L19	S38D	0.0	–0.6	–0.7	0.2	0.1	–0.6	0.7	0.2	–0.1	0.6	–0.3	2.0	Helix dipole interaction	Nicholson *et al.* (1988)
											0.6		1.5	6.9		
1L63	107L	S44G,WT*	0.0	–2.3	–0.7	–1.6	–0.6	–0.1	–0.5	–2.5	–0.53	–8.3	–1.55	3.01	Helix propensity analysis	Blaber *et al.* (1994)
											–2.62		–8.39	3.0		
1L63	1L67	L46A,WT*	1.2	–0.7	0.6	–2.5	–0.7	0.2	–2.0	–4.9	–1.86	–16.3	–6.4	6.7	Cavity creating poly Ala	Heinz *et al.* (1992)
3LZM	1L21	N55G	0.0	–2.7	–2.3	–0.4	–0.6	–0.8	0.2	–0.8	–1.5	–2.6	–1.7	2.0	Left handed helical residue	Nicholson *et al.* (1989)
											–0.6		–1.6	6.5		
1L63	1LYH	T59G,WT*	0.0	–2.4	–2.8	0.3	–0.8	0.6	–1.4	–1.8	–2.2	–6.2	–7.7	2.0	Replace helix cap	Bell *et al.* (1992)
											–1.6		–4.1	6.5		
1L63	1L76	D72P,WT*	0.0	–1.8	–0.6	–1.2	–0.7	1.0	–1.6	–3.4	–2.6	–11.4	–10.2	2.0	Proline disruption	Sauer *et al.* (1992)
											–2.7		–7.1	6.5		
3LZM	1L23	G77A	0.0	3.6	0.4	3.2	0.6	–1.1	1.7	3.8	–0.4	12.5	–1.4	2.0	Entropic stabilization	Matthews *et al.* (1987)
											0.4		0.9	6.5		
1L63	1L41	K83H,A112D,WT*	0.0	–2.0	–0.6	–1.4	0.7	0.8	–0.1	–2.0	–1.2	–6.5	–4.8	2.0	Salt bridge	Dao-pin *et al.* (1991a)
											–1.5		–3.9	6.5		
3LZM	1L48	A98V	0.9	0.9	–0.2	0.2	0.6	0.2	–0.4	–1.0	–4.9	–3.3	–14.8	3.0	Helix packing ts mutant	Dao-pin *et al.* (1991b)
3LZM	1L51	A98V,V149I,T152S	0.8	0.7	0.5	–0.7	0.6	0.3	–0.5	–1.8	4.4	6.0	–12.0	3.0	Helix packing analysis	Dao-pin *et al.* (1991b)
1L63	1L90	L99A,WT*	1.2	–0.7	1.0	–2.9	–0.7	0.4	–2.3	–5.5	–5.0	–18.4	–15.65	3.01	Cavity creating	Eriksson *et al.* (1993)
											–4.5		–11.4	5.7		
1L63	1L84′	L99A,F153A,BZ,WT*	0.0	–0.5	1.8	–2.4	–1.5	0.4	–1.9	–4.7	–7.7	–15.5	–6.9	3.01	Cavity+benzene	Eriksson *et al.* (1992b)
1L63	1L94	L99V,WT*	0.6	–0.1	0.6	–1.4	–0.1	0.3	–1.0	–2.9	–2.3	–9.7	–6.6	3.01	Hydrophobic replacement	Eriksson *et al.* (1993)
											–2.0		–5.2	5.7		
1L63	1L54	M102K,WT*	0.2	–1.6	0.0	–1.8	–0.9	2.1	–3.2	–5.5	–8.9	–18.6	–35.0	3.0	Buried lysine	Dao-pin *et al.* (1991d)
											–6.9		–20.3	5.3		

Table I. *Continued*

WT PDB code	Mutant PDB code	Mutation	$\Delta s_{jk}^{\text{Disrupt.}}$	Method						Av. est. $\Delta\hat{\Delta}G$ [*kcal*	Expt. $\Delta\Delta G$ *mol*$^{-1}$]	Av. est. $\Delta\hat{T}_m$ [℃]	Expt. ΔT_m	pH	Comment	Reference
				I			II									
				Δs_{jk}^{U}	Δ_{jk}^{F}	$\Delta\Delta s^{I}$	Δs_{jk}^{U}	Δ_{jk}^{F}	$\Delta\Delta s^{II}$							
1L63	1L77	M102L,WT*	0.4	–0.5	–1.3	0.5	0.1	–0.3	0.0	–0.3	–0.74	–1.2	–2.54	3.01	Alternative core packing	Hurley *et al.* (1992)
											–0.82		–2.31	5.7		
3LZM	1L99	Q105G	0.0	–2.8	–0.1	–2.7	–0.7	0.9	–1.5	–4.6	–1.7	–15.3	–7.2	2.1	Perturbation Trp138	Pjura *et al.* (1993a)
											–1.5		–3.9	5.8		
1L63	1L62	T109D,WT*	0.0	–1.3	–1.7	0.4	–0.1	–1.1	1.0	0.7	–0.3	2.4	–0.8	2.0	Helix dipole interaction	Nicholson *et al.* (1991)
											0.6		1.5	6.7		
3LZM	1L37	T115E	0.0	0.7	–1.9	2.6	–0.2	–1.6	1.4	2.9	–0.5	9.7	–1.7	2.0	Helix dipole/salt bridge	Dao-pin *et al.* (1991a)
											0.3		0.7	6.5		
3LZM	1L57	N116D	0.0	–0.9	–1.5	0.6	0.1	–0.7	0.8	0.6	–0.1	1.9	–0.2	2.0	Helix dipole interaction	Nicholson *et al.* (1991)
											0.6		1.6	5.7		
1L63	1TLA	S117F,WT*	1.3	1.5	–2.8	3.0	0.8	–1.0	0.5	2.4	1.4	7.7	4.8	3.0	Aromatic–aromatic interaction	Anderson *et al.* (1993)
											1.1		2.8	5.4		
1L63	118L	A130S,WT*	0.0	–1.3	0.6	–1.9	0.0	–0.1	0.0	–2.3	–1.0	–7.3	–2.89	3.0	Buried hydroxyl group	Blaber *et al.* (1993)
3LZM	1L33	V131A	0.0	–0.9	–1.9	1.0	–0.6	–1.2	0.6	0.8	0.15	2.7	0.6	2.03	Strain release	Dao-pin *et al.* (1990)
											0.3		0.9	2.83		
3LZM	1L70	V131A,N132A	0.0	0.1	–1.9	2.1	–0.6	–1.7	1.1	2.2	0.62	7.3	2.3	2.0	Poly Ala helix 126–134	Zhang *et al.* (1991)
3LZM	1L69	L133A	1.2	–0.7	1.0	–2.9	–0.7	0.4	–2.3	–5.5	–4.2	–18.4	–17.1	2.0	Cavity creating poly Ala	Eriksson *et al.* (1992a)
											–3.6		–10.55	3.01		
3LZM	1L20	N144D	0.0	–0.9	–1.8	0.8	0.1	–0.6	0.7	0.7	–0.1	2.5	–0.2	2.0	Helix dipole interaction	Nicholson *et al.* (1988)
											0.5		1.4	6.9		
1L63	1L85	F153A,WT*	1.3	0.2	1.1	–2.3	–0.7	0.0	–2.1	–4.8	–3.5	–16.0	–12.3	3.01	Cavity creating	Eriksson *et al.* (1993)
											–3.8		–9.3	5.7		
2LZM	1L16	G156D	1.1	1.8	3.5	–2.9	0.7	0.8	–1.3	–4.5	–1.4	–14.8	–5.1	2.0	Random ts mutant	Gray and Matthews (1987)
											–2.3		–6.1	6.5		
2LZM	1L06	T157E	0.0	0.7	2.0	–1.3	–0.2	0.4	–0.6	–2.4	–1.5	–8.0	–5.8	2.0	H-bond analysis	Alber *et al.* (1987)

Component terms used to calculate $\Delta\Delta s$ according to Equation 6 are indicated. Method I is based on structural environment-dependent amino acid substitution patterns, and Method II utilizes environmental amino acid propensity tables. Free energy difference estimates ($\Delta\Delta G^{\text{I}}$ and $\Delta\Delta G^{\text{II}}$) can be calculated from best-fit slope and ordinate intercept regression parameters of plots of $\Delta\Delta s^{\text{I}}$ and $\Delta\Delta s^{\text{II}}$ versus $\Delta\Delta G$ (see legends to Figures 2b and 3b). The average of these estimates (Av. $\Delta\Delta G$) shown in the table may be compared directly with experimentally observed values of $\Delta\Delta G$. Average estimates of ΔT_{m} were derived in an analogous way. WT* denotes a pseudo-wild-type T4 lysozyme in which the two cysteines present in the normal wild-type enzyme are replaced with threonine and alanine (C54T,C97A).

†The Brookhaven Data Bank coordinate set, 3LZM was used as the wild type, because the atomic positions for Leu164 are absent in 1L63.

‡Measurement in the presence of 10.4 mM benzene; disruption term ($\Delta s_{jk}^{\text{Disrupt}}$) is not used.

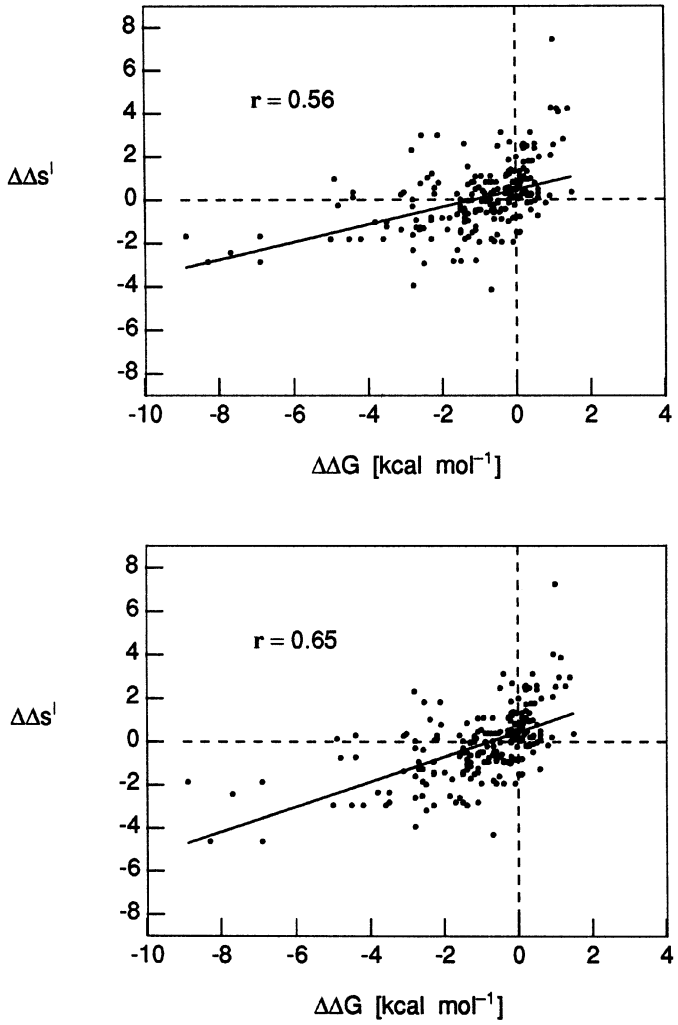

Fig. 2. Calculated differences in stability scores ($\Delta\Delta s^{I}$) versus experimentally observed free energy differences ($\Delta\Delta G$) between wild-type and mutant T4 lysozyme proteins. Values of $\Delta\Delta s^{I}$ were calculated using Method I, based on structural environment-dependent amino acid substitution patterns, according to (**a**) Equation 3 or (**b**) Equation 6, incorporating the disruption penalty term (Equation 7) at buried residue positions. A total of 217 points are plotted, corresponding to 159 mutants for which experimental values of $\Delta\Delta G$ were available in the literature. These are detailed in the anonymous ftp site, together with the PDB codes of the wild-type and mutant coordinate sets used to determine the residue environments. Robust weighted regression least-squares fitting of the data gave correlation coefficients (r) of (a) 0.56 and (b) 0.65. The corresponding unweighted linear regression correlation coefficients are 0.36 and 0.52, respectively. The best-fit regression line in (a) has a slope of 0.40 ± 0.04 kcal/mol and an ordinate intercept of 0.46 ± 0.08. The regression line in (b) has a slope of 0.58 ± 0.05 kcal/mol and an ordinate intercept of 0.43 ± 0.09.

Fig. 3a

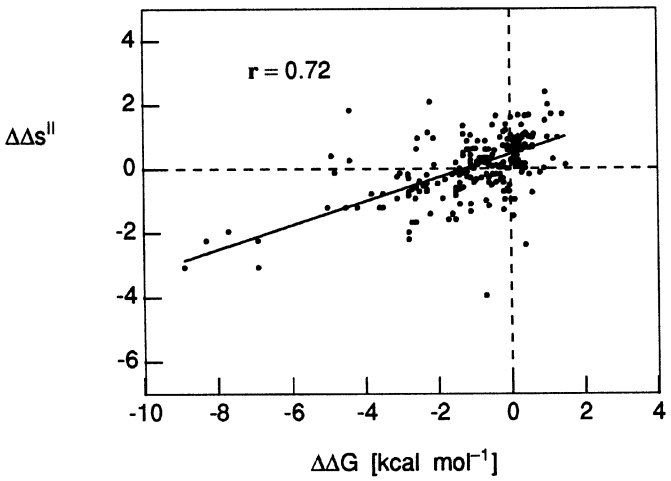

Fig. 3b

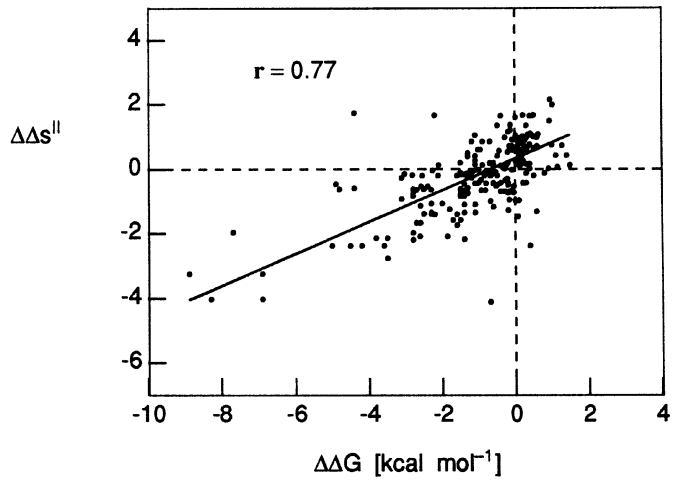

Fig. 3. Calculated differences in stability scores ($\Delta\Delta s^{II}$) versus experimentally observed free energy differences ($\Delta\Delta G$) between wild-type and mutant T4 lysozyme proteins. Values of $\Delta\Delta s^{II}$ were calculated using Method II, based on structural environment-dependent amino acid propensities according to (**a**) Equation 3 or (**b**) Equation 6, incorporating the disruption penalty term (Equation 7) at buried residue positions. Full details of all the mutations, the coordinate data sets used and the experimental values of $\Delta\Delta G$ are given in the anonymous ftp site and correspond to those in Figure 2. Robust weighted regression least-squares fitting of the data yielded correlation coefficients (r) of (a) 0.72 and (b) 0.77. The corresponding unweighted linear regression correlation coefficients are 0.44 and 0.64 respectively. The best-fit regression line in (a) has a slope of 0.37 ± 0.03 kcal/mol and an ordinate intercept of 0.46 ± 0.05. The regression line in (b) has a slope of 0.49 ± 0.03 kcal/mol and an ordinate intercept of 0.33 ± 0.05.

contribution of $\Delta s_{jk}^{Disrupt}$ to $\Delta\Delta s$ is 41% in the case of Method I and 52% for Method II.

The disruption term is also applied in cases such as S117F, in which there is an increase in size in the buried mutant side chain. The T4 lysozyme mutant S117F was isolated fortuitously by Anderson *et al.* (1993) and was found to be more thermostable than the wild type by 1.1–1.4 kcal/mol. Figure 5 shows that the burial of Phe117 can be accommodated in part by the presence of an internal cavity adjacent to Leu133, but also by rearrangements of the surrounding side chains of Leu121, Leu133 and Phe153 and shifts in the main chain. Together these result in a minimal increase in packing density. The energetic cost of structural rearrangement appears to be offset by the energy gain from the burial of Phe117 plus its edge–edge interaction with Phe153 and the loss of a destabilizing, strained hydrogen bond formed between the γ-hydroxyl of Ser117 and the side chain of Asn132 (Blaber *et al.*, 1995). The S117F mutant is correctly predicted by Methods I and II to be more stable, the disruption penalty term being offset by positive values of $\Delta s^{U}_{jk} - \Delta s^{F}_{jk}$ (see Table I). Also worth noting is the successful prediction of positive values for the N132A-containing double mutant V131A,N132A (Zhang *et al.*, 1991), in which the buried hydrogen bond is also lost (Table I).

Table II. Performance of thermal stability prediction algorithm: T4 lysozyme mutants

Plot versus $\Delta\Delta G$			Statistic	Residue environment						
Figure	$\Delta s_{jk}^{Disrupt}$	Dependent variable		Hydrogen bond(s)			Relative side-chain accessibility		Secondary structure	
				All	√	×	<7%	⩾7%	Helix/sheet	Coil
2a	×	$\Delta\Delta s^{I}$	$t > 1.96$ (%)	6.6	11.1	3.3	7.2	5.4	6.9	3.2
			$\Delta\Delta G$ ± agreement (%)	63.1	52.1	69.1	69.7	60.1	65.8	54.5
2b	√	$\Delta\Delta s^{I}$	$t > 1.96$ (%)	4.6	7.1	3.3	4.3	4.8	5.7	1.2
			$\Delta\Delta G$ ± agreement (%)	65.4	53.2	72.0	75.9	60.6	68.8	54.5
3a	×	$\Delta\Delta s^{II}$	$t > 1.96$ (%)	3.7	5.1	2.9	5.8	2.7	4.5	1.0
			$\Delta\Delta G$ ± agreement (%)	65.4	64.0	66.2	62.4	66.9	66.1	63.2
3b	√	$\Delta\Delta s^{II}$	$t > 1.96$ (%)	3.7	4.4	3.3	8.0	1.7	4.5	1.0
			$\Delta\Delta G$ ± agreement (%)	73.2	68.0	76.2	83.7	68.5	76.4	63.2
6a	√	$\Delta\Delta s^{I}$	$t > 1.96$ (%)	5.1	6.3	4.4	2.6	6.2	6.3	1.2
			$\Delta\Delta G$ ± agreement (%)	66.4	53.2	73.4	78.8	60.6	68.9	58.4
6b	√	$\Delta\Delta s^{II}$	$t > 1.96$ (%)	6.5	1.8	9.0	13.9	3.0	6.3	6.8
			$\Delta\Delta G$ ± agreement (%)	74.6	66.6	78.9	86.6	69.2	78.8	61.2
7a	√	av. est. $\Delta\hat{\Delta}G$	$t > 1.96$ (%)	4.1	7.3	2.4	6.1	3.3	5.1	1.2
			$\Delta\hat{\Delta}G$ ± agreement (%)	77.4	76.2	78.1	89.4	71.9	76.7	79.7
7b	√	av. est. $\Delta\hat{\Delta}G$	$t > 1.96$ (%)	6.5	1.8	9.0	13.9	3.0	6.3	6.8
			$\Delta\hat{\Delta}G$ ± agreement (%)	77.4	73.8	79.4	90.5	71.4	78.5	73.9

The fraction of residuals from the least-squares regression analyses having t values >1.96, and the fraction of predictions correctly categorized as either stabilizing or destabilizing, are given for individual plots as indicated. All the statistical analyses are based on 217 data points. Point mutation sites are subdivided according to wild-type residue environment. The reported percentages refer to the fraction of data belonging to the same environmental class. In the case of multiple-site mutants, the residue environment at each point mutation site is considered to contribute equally to a single prediction. This accounts for apparently anomalous changes in the percentage of predictions involving mutations at partially buried and surface-exposed sites correctly categorized as stabilizing or destabilizing upon introduction of the disruption term at buried mutation sites.

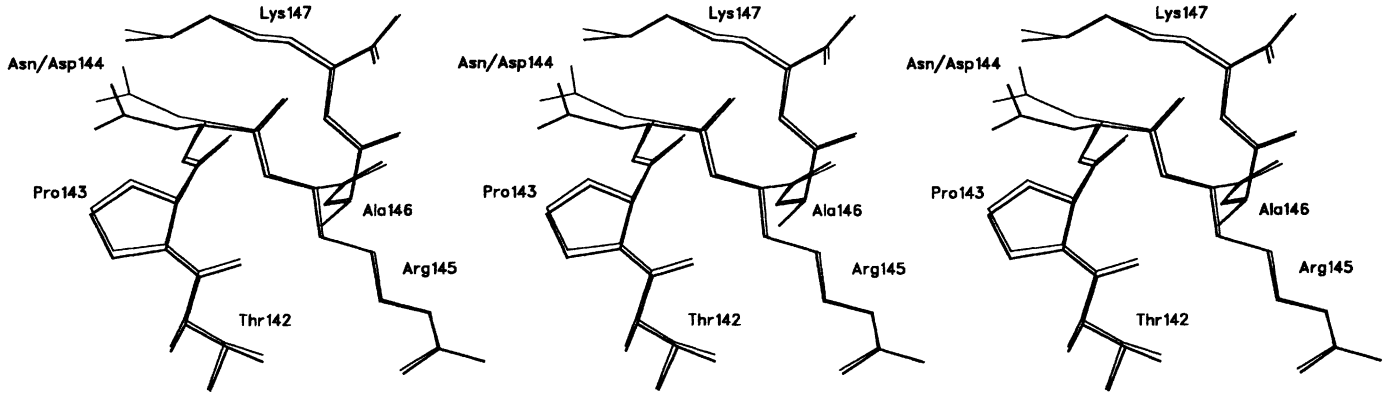

Fig. 4. Superposition of N144D (PDB code 1L20) T4 lysozyme and wild type (PDB code 3LZM). Superposition of the mutant (thick bonds) on the wild type (thin bonds) is based on the optimum alignment of main-chain (N, C^{α} and C) atoms of the C-terminal domains (residues 80–160; r.m.s. deviation = 0.07 Å). The engineered aspartic acid side chain stabilizes the mutant structure by 0.6 kcal/mol relative to the wild type when ionized at pH 6.7 (Nicholson *et al.*, 1988). Stabilization results from a generalized electrostatic interaction of the carboxylate side chain and the positive charge at the N-terminus of helix I (residues 137–141) rather than hydrogen bond formation with the peptide nitrogens within the first turn of the helix. Similar conformations are observed in the T109D and N116D mutant structures that display favourable Ncap +1/α-helix dipole interactions. The triptic stereoview was prepared using the SETORPLOT extension facility of the SETOR molecular graphics display program (Evans, 1993).

No attempt was made to refine the disruption term by relating it more closely to the actual change in cavity volume calculated from the wild-type and mutant protein atomic coordinates. In implementations of the method (see below), mutant coordinate sets are either not used for operational reasons of computational efficiency or are modelled from the wild type, in which case they are not likely to be sufficiently reliable, particularly if the backbone is kept fixed, to determine the cavity volume change with any accuracy. The present form of the disruption term was therefore retained in all further tests of the method.

Whilst the results obtained using the mutant atomic coordinate sets correlate well with the experimental thermodynamic data in general, it is important to apply tests of performance under more realistic conditions in which structural information is only available for the wild-type protein. A second set of predictions, based on a description of the mutant environmental features determined by extrapolation from wild-type T4 lysozyme, was therefore made (Figure 6 and Table II). The extrapolation procedure retains the notion of a fixed backbone, inherent in established inverse folding protocols in which different sequences are threaded over a given fold with the aim of identifying sequence structure matches of homologous proteins (see, for example, Bowie *et al.*, 1991; Jones *et al.*, 1992; Johnson *et al.*, 1993). Hydrogen bonding combinations are, however, adjusted to avoid incompatibilities in residue

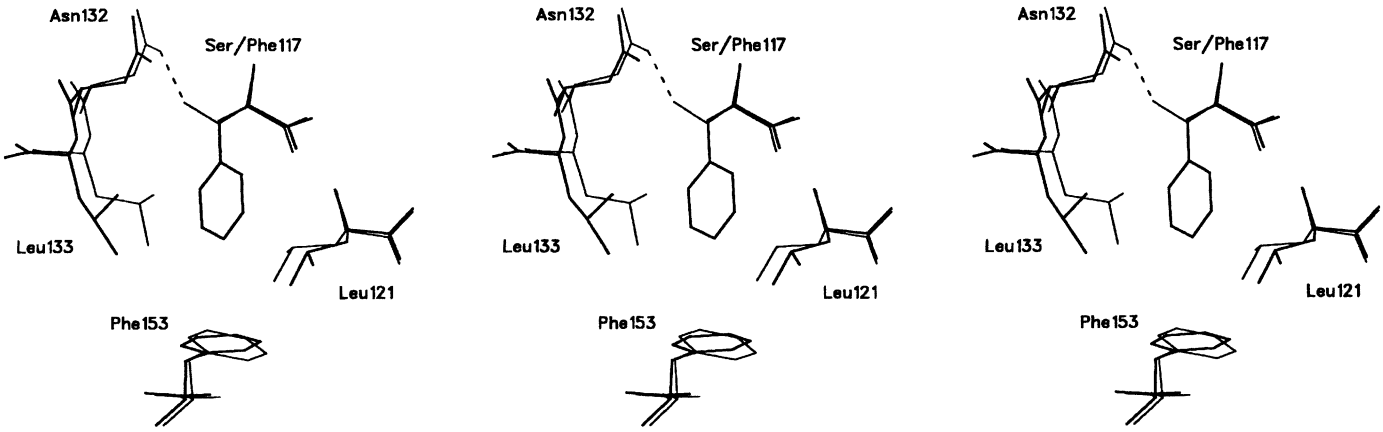

Fig. 5. Triptic stereoview of the superposition of mutant S117F (PDB code 1TLA) T4 lysozyme and C54T,C97A pseudo-wild-type (PDB code 1L63) structures. Superposition of the thermostable mutant (thick bonds) on the wild type (thin bonds) was carried out using the main-chain (N, C^{α} and C) atoms of the C-terminal domains (residues 80–160). The r.m.s. deviation was 0.20 Å. Burial of Phe117 is accompanied by hydrophobic core repacking involving rearrangement of the Leu121, Leu133 and Phe153 side chains and main-chain shifts. The χ^1 side-chain torsion angle is rotated by ~150° in the Phe117 mutant relative to the wild type, facilitating a near optimal aromatic ring edge–face interaction with Phe153. A short hydrogen bond (2.34 Å) formed in the wild type between O^{γ} of Ser117 and Asn132, depicted as a broken line, is also lost in the S117F mutant structure. The diagram was prepared using the SETORPLOT and SETOR programs (Evans, 1993).

type. The *r* values for the robust regression analyses shown in Figure 6 are 0.65 and 0.74 for Methods I and II, respectively, and are similar to values of 0.65 and 0.77 obtained using crystallographic data. The percentage of points having *t* values >1.96 is seen to increase from 3.7 to 5.1% for Method I, and to 6.5% for Method II (Table II). When performance is assessed on the basis of the correct assignment of destabilizing or stabilizing residue replacements, the environment extrapolation procedure closely reproduces the results obtained for the crystal structures, with Methods I and II showing 66.4 and 74.6% agreement with the experimental data, respectively (Table II).

Although stability difference scores generated using the alternative methods are derived from statistical analyses of the same multi-body interactions codified in the protein structural alignment database, the different information content of the substitution and propensity tables (see Topham *et al.*, 1993) does not permit the combination of $\Delta\Delta s^{I}$ and $\Delta\Delta s^{II}$. Thus, to average the results it is first necessary to transform the two sets of stability difference scores into estimates of $\Delta\Delta G$ ($\Delta\hat{\Delta}G$). This was carried out using the slope and ordinate intercept values determined from the least-squares regression analyses of $\Delta\Delta s$ on experimentally determined values of $\Delta\Delta G$ (see the legends to Figures 2 and 3). Calculated average values of $\Delta\hat{\Delta}G$ for the mutant T4 lysozyme crystal structures are plotted in Figure 7a, from which a regression coefficient of 0.74 was determined. As expected, the slope of the plot is not significantly different from unity (1.01), and the best-fit regression line passes close to the origin. Agreement with the experimental data in terms of correct sign assignment of the stability difference scores is 77.4% for all the data and 89.4% for mutations at buried residue positions (Table II). When the procedure is applied to stability difference scores derived using mutant residue environments extrapolated from the wild type, in conjunction with slope and intercept estimates from the mutant crystal structure data sets, $r = 0.72$ (Figure 7b). The slope of the regression line is again close to unity (0.97 ± 0.06) and the ordinate intercept is –0.12 ± 0.12 kcal/mol. The percentage sign agreement (77.4%, Table II) is identical to that obtained using the actual crystal structure environments.

Correlations of the stability difference scores are also found with differences in melting temperature. Plots (data not shown) of $\Delta\Delta s^{I}$ and $\Delta\Delta s^{II}$ versus experimentally observed values of ΔT_m yield values of 0.58 and 0.68, respectively. The best-fit values of the slopes and ordinate intercepts of these plots were then used to calculate two sets of estimates ($\Delta\hat{T}^{I}_{m}$ and $\Delta\hat{T}^{II}_{m}$) of the change in melting temperature for each mutant. When these are averaged and subjected to a regression analysis, a correlation coefficient of 0.65 is obtained.

Predictions of barnase staphylococcal nuclease mutant stability based on wild-type structures

Further tests of the thermal stability predictions were then carried out on 68 barnase and 83 staphylococcal nuclease mutants using only structural information from the wild-type protein crystal structures. Taken as a single set, 34.8% of the mutations occur at buried sites, 30.8% are involved in side-chain hydrogen bond formation and 32.3% of mutated residues lie outside secondary structural regions. Results for individual mutants based on the determination of mutant residue environments by extrapolation from wild-type coordinate sets are reported in the anonymous ftp site. Regression analyses of the plotted data (Figure 8) again indicate significant correlation with experimental measurements of $\Delta\Delta G$. Correlation coefficients of 0.76 were calculated for Method I, 0.79 for Method II and 0.80 when the $\Delta\hat{\Delta}G^{I}$ and $\Delta\hat{\Delta}G^{II}$ estimates are averaged. Linear transformation of $\Delta\Delta s^{I}$ and $\Delta\Delta s^{II}$ to yield the estimates $\Delta\hat{\Delta}G^{I}$ and $\Delta\hat{\Delta}G^{II}$ was carried out using slope and ordinate intercept estimates from fits of the mutant T4 lysozyme stability difference scores, similarly calculated on the basis of extrapolated residue environments, on $\Delta\Delta G$ (see the legend to Figure 6). The slopes of the regression lines shown in Figure 8a and b for Method I (1.20 ± 0.08) and Method II (0.98 ± 0.06) are both close to an idealized value of unity. However, the ordinate intercepts are underestimated by ~1 kcal/mol (Method I, –0.82 ± 0.27 kcal/mol; Method II, –1.03 ± 0.20 kcal/mol). Intermediate values of 1.09 ± 0.07 for the slope and –0.92 ± 0.22 kcal/mol are obtained for the averaged $\Delta\hat{\Delta}G$ data (Figure 8c). Given that the majority (95%) of mutations

Fig. 6a

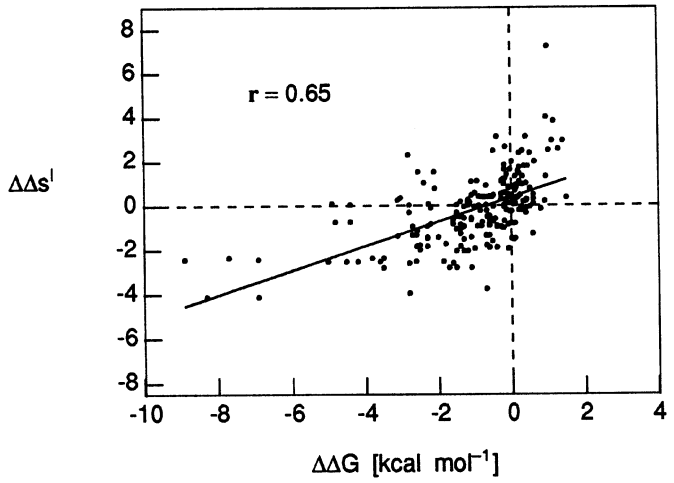

Fig. 6b

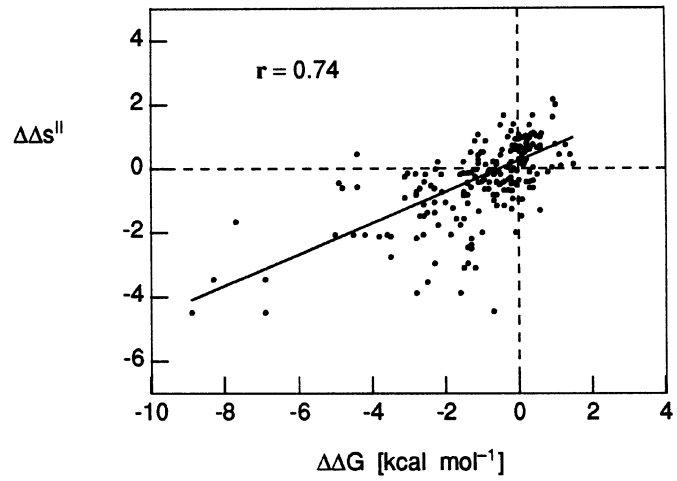

Fig. 6. Stability difference scores for mutation ($\Delta\Delta s$) based on residue environment extrapolation from wild-type T4 lysozyme versus experimentally observed free energy differences ($\Delta\Delta G$). Values of $\Delta\Delta s$ were calculated using (**a**) Method I and (**b**) Method II according to Equation 6. The mutations correspond to those of Figures 2 and 3. The physical environments of mutated residues were determined as described in Materials and methods using wild-type PDB coordinate sets. Robust weighted regression least-squares fitting of the data gave correlation coefficients *r*() of (a) 0.65 and (b) 0.74. The corresponding unweighted linear regression correlation coefficients are 0.54 and 0.67 respectively. The best-fit regression line in (a) has a slope of 0.55 ± 0.04 kcal/mol and an ordinate intercept of 0.36 ± 0.08. The regression line in (b) has a slope of 0.49 ± 0.03 kcal/mol and an ordinate intercept of 0.24 ± 0.06.

Fig. 7a

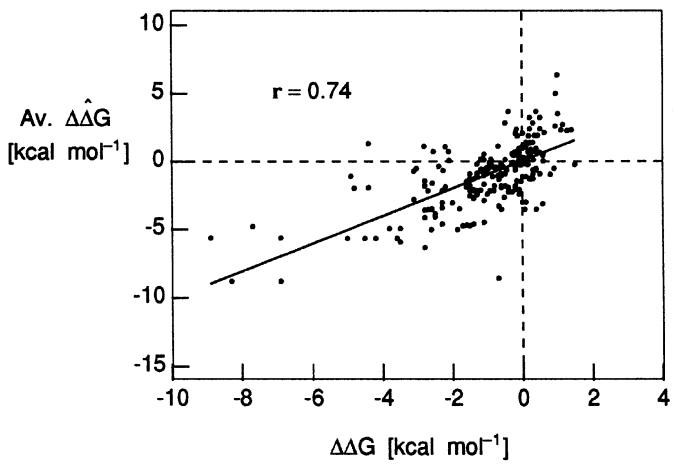

Fig. 7b

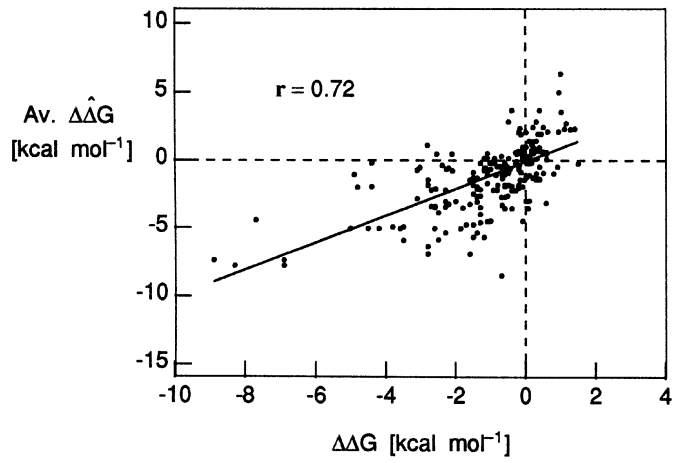

Fig. 7. Average of Method I and II estimates of stability (Av. $\Delta\hat{\Delta}G$) versus experimentally observed values of $\Delta\Delta G$ for T4 lysozyme mutants. Calculations of Av. $\Delta\hat{\Delta}G$ were based on (**a**) crystal structure environments and (**b**) residue environment extrapolation from wild-type PDB coordinate sets. Values of Av. $\Delta\Delta G$ were calculated from estimates of $\Delta\hat{\Delta}G^{I}$ and $\Delta\hat{\Delta}G^{II}$, obtained by linear transformation of $\Delta\Delta s^{I}$ and $\Delta\Delta s^{II}$ using the best-fit slope and intercept estimates given in Figures 2b and 3b, respectively. Av. estimated $\Delta\hat{\Delta}G$ values in (a) and the corresponding experimental values of $\Delta\Delta G$ for individual mutants are tabulated in our ftp site. Robust weighted regression least-squares fitting of the data gave correlation coefficients (*r*) of (a) 0.74 and (b) 0.72. The corresponding unweighted linear regression correlation coefficients are 0.71 and 0.72, respectively. The best-fit regression line in (a) has a slope of 1.01 ± 0.06 and an ordinate intercept of 0.02 ± 0.12 kcal/mol. The regression line in (b) has a slope of 1.00 ± 0.07 and an ordinate intercept of –0.12 ± 0.12 kcal/mol.

are destabilizing, an average figure of 85.8% of predictions correctly classified as either destabilizing or stabilizing, based on the untransformed stability difference score data in Table III ($\Delta\Delta s^{I}$, 84.8%; $\Delta\Delta s^{II}$, 86.8%), provides a more reliable indicator of predictive power than does the corresponding average $\Delta\hat{\Delta}G$ value (90.1%, Table III). Table III shows that mutations at buried residue positions are the best predicted (Method I, 92.4% agreement; Method II, 98.1%), whilst predictions for residues involved in hydrogen bonding are in least agreement (72.0 and 78.5% for Methods I and II, respectively).

Predictions for the T99V barnase mutant and for four other barnase mutants involving the replacement of Tyr17 by Ala (Y17A; Y13A,Y17A; T16A,Y17A; T16S,Y17A) are associated with *t* values >1.96 using either Method I or II (see Figure 8). Although these mutants are destabilizing with experimentally determined estimates of $\Delta\Delta G$ lying in the range –2.0 to –4.2 kcal/mol, all of the outlying values of $\Delta\Delta s^{I}$ and $\Delta\Delta s^{II}$ are erroneously calculated to be ⩾0. Thr99 is a buried residue in β-sheet 4 with its side chain involved in hydrogen bonds (Figure 9) with the main-chain Tyr103 –NH group, the main-chain carbonyl of Thr105 and the side chain of Asp101. Isosteric mutation to valine removes these hydrogen bonds, whilst producing a minimal change in residue volume and

side-chain geometry, thus maintaining van der Waals contacts. As there is insufficient room for a water molecule, the stabilization energy gained by burying a hydrophobic group instead of a hydrophilic group is opposed by the destabilizing effect of unpairing of the hydrogen bond donor and acceptor groups of the threonine side-chain –OH, leading to a net loss in stability of 2.67 kcal/mol (Serrano *et al.*, 1992). The application of Method I or II results in an apparent decrease in stability of the unfolded state upon the replacement of threonine by valine, manifested by positive values of Δs_{jk}^{U} that favour an increase in the overall stability. However, in neither application is this offset by a greater decrease in the stability of the folded state represented by the combination of Δs_{jk}^{F} and an appropriately small disruption term. This is explained in the case of Method I by an adverse probability ratio of 2.3 for the substitution of a buried hydrogen bonded threonine by valine (in any environment) compared with the substitution of a buried valine by threonine (in any environment). The loss of the hydrogen bond with the main-chain –NH group is likely to be the most important of the three types considered. However, in calculating the probability of the replacement of threonine by valine, for reasons of database size limitation, we have used substitution tables obtained by averaging over the different hydrogen bonding combinations, resulting in a higher substitution probability. Conversely, an underestimation of the likelihood of the reverse substitution of valine to threonine is probably a consequence of not taking account of the environment of threonine residue environment, again for reasons of limited data availability.

Tyr17 is a surface-exposed residue in α-helix 1 of barnase (Serrano *et al.*, 1990). The resistance of the Y17A-containing mutant quartet to prediction stems in the case of Method I from the contribution of large positive values of Δs_{jk}^{U}. This results from a tendency for tyrosine residues in surface-accessible coil regions used to model the unfolded state to be evolutionarily conserved to a relatively high degree, possibly because naturally conservative mutations to phenylalanine are less likely in these regions. In the case of Method II, negative values of Δs_{jk}^{F} provide the dominant contribution to $\Delta\Delta s^{II}$, reflecting the greater helix-forming propensity (×3.5) of alanine compared with that of non-hydrogen bonded tyrosine at exposed helix faces. The probability ratio is approximately halved (1.7) when alanine and (non-hydrogen bonded) tyrosine occur at partially buried positions. This is sufficient to give a negative value of $\Delta\Delta s^{II}$ for the Y13A barnase mutant.

Calculations of predicted barnase and staphylococcal nuclease mutant thermal stabilities were also made using model coordinate sets in which the backbone was fixed and the side chains swapped (Figure 10 and Table III). The results are in close general agreement with those based on the determination of mutant residue environment by extrapolation from the wild

Fig. 8a

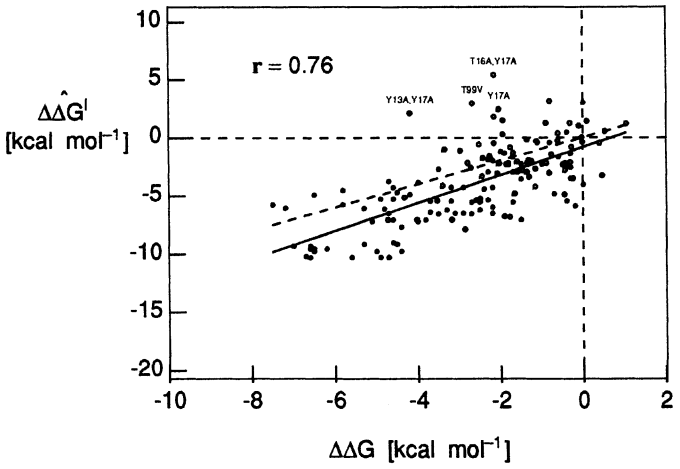

Fig. 8b

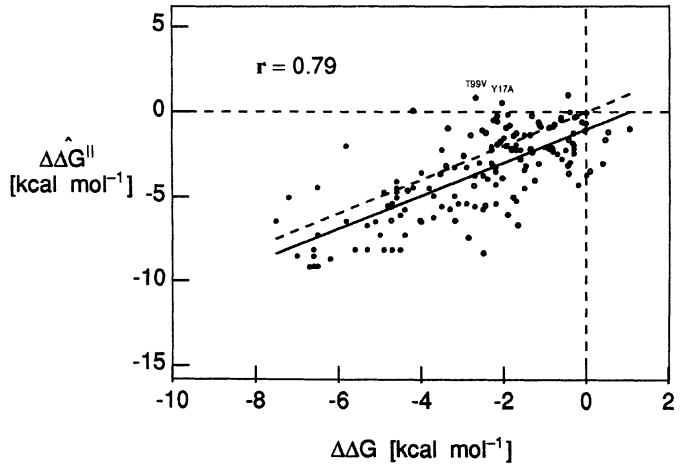

Fig. 8c

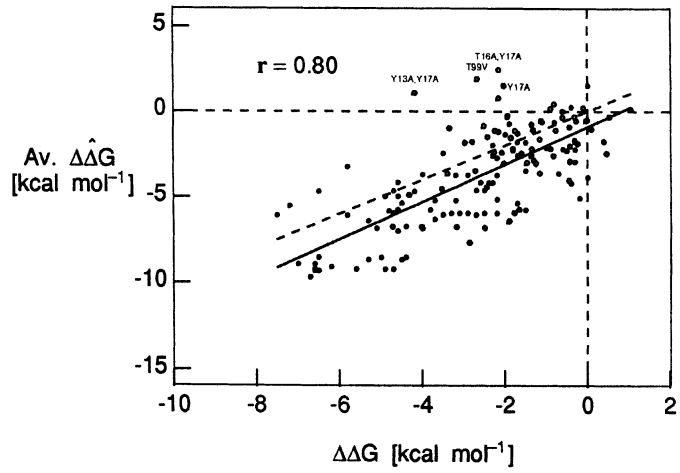

Fig. 8. Calculated stabilities ($\Delta\Delta\hat{G}$) versus experimentally observed values of $\Delta\Delta G$ for 83 staphylococcal nuclease and 68 barnase mutants. Calculations of (**a**) $\Delta\Delta\hat{G}^{I}$ using propensity tables (Method I), (**b**) $\Delta\Delta\hat{G}^{II}$ using substitution tables (Method II), and (**c**) their average (Av. $\Delta\Delta\hat{G}$) were based on residue environment extrapolation from the staphylococcal nuclease (1STN) and barnase (1BGS, A chain) wild-type PDB coordinate sets. Stability estimates for staphylococcal nuclease (●) and barnase (○) and the corresponding experimentally observed values for individual mutants are listed in the anonymous ftp site. Estimates of $\Delta\Delta\hat{G}^{I}$ and $\Delta\Delta\hat{G}^{II}$ were obtained by linear transformation of $\Delta\Delta s^{I}$ and $\Delta\Delta s^{II}$, respectively, using the best-fit slope and intercept estimates determined for the T4 lysozyme mutants based on residue environment extrapolation (see the legend to Figure 6). Robust weighted regression least-squares fitting of the data gave correlation coefficients (r) of (a) 0.76, (b) 0.79 and (c) 0.80. The corresponding unweighted linear regression correlation coefficients are 0.86, 0.90 and 0.89, respectively. The best-fit regression (solid) line parameters are: (a) slope = 1.20 ± 0.08, ordinate intercept = –0.82 ± 0.27 kcal/mol; (b) slope = 0.98 ± 0.06, ordinate intercept = –1.03 ± 0.20 kcal/mol; (c) slope = 1.10 ± 0.07, ordinate intercept = –0.92 ± 0.22 kcal/mol. The idealized dashed lines depicted in each plot have unit slope and pass through the origin.

Table III. Performance of thermal stability prediction algorithm: staphylococcal nuclease and barnase mutants

Plot versus $\Delta\Delta G$		Statistic	Residue environment						
Figure	Dependent variable		Hydrogen bond(s)			Relative side-chain accessibility		Secondary structure	
			All	√	×	<7%	⩾7%	Helix/sheet	Coil
8a	$\Delta\hat{\Delta}G^{I}$	$t > 1.96$ (%)	4.0	6.5	2.9	1.9	5.1	5.9	0.0
		$\Delta\Delta G$ ± agreement (%)	88.1	74.2	94.3	94.3	84.8	88.7	86.7
	$\Delta\Delta s^{I}$	$\Delta\Delta G$ ± agreement (%)	84.8	72.0	90.4	92.4	80.7	85.3	83.6
8b	$\Delta\hat{\Delta}G^{II}$	$t > 1.96$ (%)	4.6	4.3	4.8	5.7	4.1	4.9	4.1
		$\Delta\Delta G$ ± agreement (%)	92.1	84.9	95.2	98.1	88.8	92.0	88.1
	$\Delta\Delta s^{II}$	$\Delta\Delta G$ ± agreement (%)	86.8	77.4	90.9	98.1	80.7	87.2	81.9
8c	av. est. $\Delta\hat{\Delta}G$	$t > 1.96$ (%)	4.6	4.3	4.8	1.9	6.1	4.9	4.1
		$\Delta\Delta G$ ± agreement (%)	90.1	78.5	95.2	98.1	85.8	90.0	90.1
10a	$\Delta\hat{\Delta}G^{I}$	$t > 1.96$ (%)	4.6	4.3	4.8	5.7	4.1	6.9	0.0
		$\Delta\Delta G$ ± agreement (%)	87.4	74.2	93.3	94.3	83.8	87.8	86.7
	$\Delta\Delta s^{I}$	$\Delta\Delta G$ ± agreement (%)	83.5	72.0	88.5	92.4	78.7	83.4	83.6
10b	$\Delta\hat{\Delta}G^{II}$	$t > 1.96$ (%)	4.0	4.3	3.8	3.8	4.1	3.9	4.1
		$\Delta\Delta G$ ± agreement (%)	91.4	84.9	94.3	98.1	87.8	93.0	88.1
	$\Delta\Delta s^{II}$	$\Delta\Delta G$ ± agreement (%)	87.4	79.6	90.9	98.1	81.7	90.0	81.9
10c	av. est. $\Delta\hat{\Delta}G$	$t > 1.96$ (%)	4.6	4.3	4.8	3.8	5.1	5.9	2.0
		$\Delta\Delta G$ ± agreement (%)	90.1	78.5	95.2	98.1	85.8	90.0	90.1

The fraction of residuals from the least-squares regression analyses having t values >1.96, and the fraction of predictions correctly categorized as either stabilizing or destabilizing, are recorded for each plot. As $\Delta\Delta G$ values are obtained by linear transformation of $\Delta\Delta s$, plots of either variable versus $\Delta\Delta G$ are characterized by a common regression coefficient and yield identical values of t. The statistical analyses are all based on a combined set of 151 data points. Point mutation sites may be subdivided according to wild-type residue environment. The percentages refer to the fraction of data belonging to the same environmental class. For multiple-site mutants, the residue environment at each point mutation site is considered to contribute equally to a single prediction.

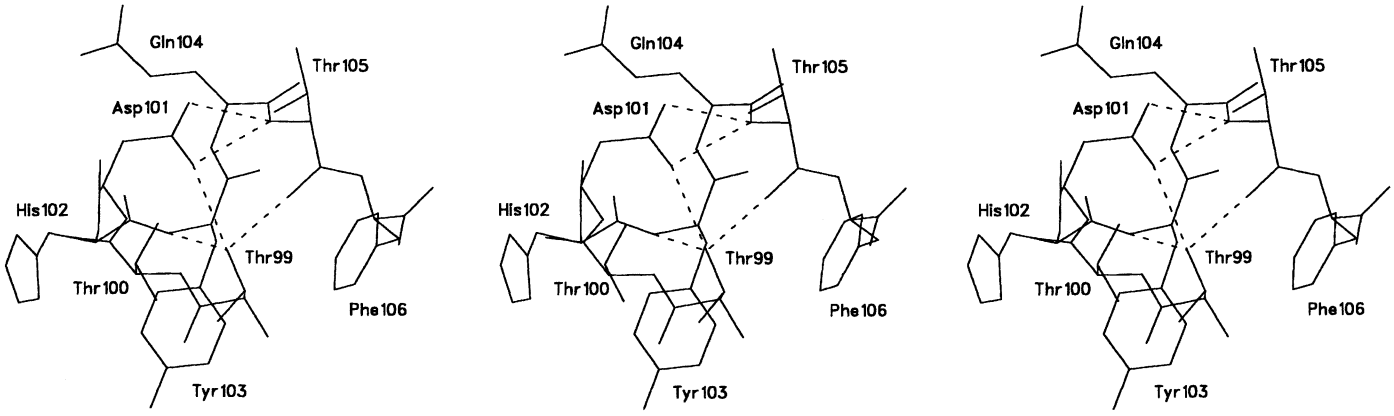

Fig. 9. View of some of the interactions in the structural environment surrounding Thr99 in wild-type barnase (PDB code 1BGS A chain). Hydrogen bonds determined using the distance criterion of Overington *et al.* (1990) are drawn as broken lines. The triptic stereoview was prepared using the SETORPLOT extension facility of the SETOR molecular graphics display program (Evans, 1993).

type. Correlation coefficients of 0.75 were calculated for Method I, 0.77 for Method II and 0.78 when the $\Delta\hat{\Delta}s^{I}$ and $\Delta\hat{\Delta}s^{II}$ estimates are averaged. The statistics for the number of outlying data points with t values >1.96 and the fraction of predictions correctly classified as stabilizing or destabilizing also remain essentially unchanged. These findings are important from the point of view of implementation of the thermal stability prediction algorithm because the extrapolation procedure is more amenable to the rapid evaluation of amino acid replacement at any position in a protein fold. A limitation of the extrapolation protocol is the neglect of putative hydrogen bonds in mutants resulting from non-polar/hydrophobic → polar/charged residue replacements. In the screening of a large number of potential mutants, it is therefore necessary to allow for such 'phantom' hydrogen bonds by making more than one prediction. A two-stage process can thus be envisaged. Model atomic coordinate sets of the more promising candidates identified from an initial screening process, based on environment extrapolation, could be built and their thermal stabilities re-assessed in a second screening phase. De Fillipis *et al.* (1994) have recently proposed a method for the determination of the local structural environment of point-site mutants based on database searches of statistically preferred and sterically admissible side-chain rotomers.

Conclusions

The overall results of the predictions demonstrate a good correlation between the predicted and experimentally observed $\Delta\Delta G$ and ΔT_m values for a wide variety of mutations of T4 lysozyme despite the dependency of melting temperature and, to a lesser extent, $\Delta\Delta G$ on the pH (Yang and Honig, 1993) and solvent conditions. The correlation with experimentally

observed $\Delta\Delta G$ values was not unduly adversely affected when the predictions were made without the input of structural information provided by the mutant crystal structures. Similar correlations were obtained in tests using a combined set of 151 barnase and staphylococcal nuclease mutants. The underlying model which caters for both the unfolded and folded states would therefore appear to provide coarse, but arguably useful, predictions of relative thermal stability at very little computational cost. This should prove valuable in the evaluation of large numbers of possible mutants prior to mutagenesis experiments. As van Gunsteren and Mark (1992) have pointed out, the accuracy of any model will remain limited by the crudest approximation or the weakest assumption. In this case the use of a single substitution or propensity table to represent the unfolded state is probably a significant limiting factor. In common with the findings of Ota *et al.* (1995), we report that mutations occurring at buried sites are more reliably predicted than mutations at exposed or partially buried positions. Another emergent trend is the greater success achieved for residues not engaged in hydrogen bonding interactions with neighbouring residues compared with those that are. This may be a consequence in the case of the substitution table method of both averaging data over the different hydrogen bonding classes and not having sufficient data to permit the calculation of genuinely multi-dimensional tables, in which the hydrogen bonding class of the substituting residue is considered. Neglect of the environmental class of the substituting residue may also provide an explanation for the somewhat marginally better performance of the propensity table-based method, which was most marked for the T4 lysozyme mutant set. Optimum choice of physical features describing native protein structure is also important, as demonstrated by Gilis and Rooman (1996). The features used here and elsewhere, to rank loop fragments in modelling protein structure (Topham *et al.*, 1993), to probe local protein structure integrity (Topham *et al.*, 1994) or in the identification of sequences that adopt the same fold as a known structure (Johnson *et al.*, 1993), are all local in nature. Longer range electrostatic contributions, for example, could be parameterized for use in substitution and propensity tables.

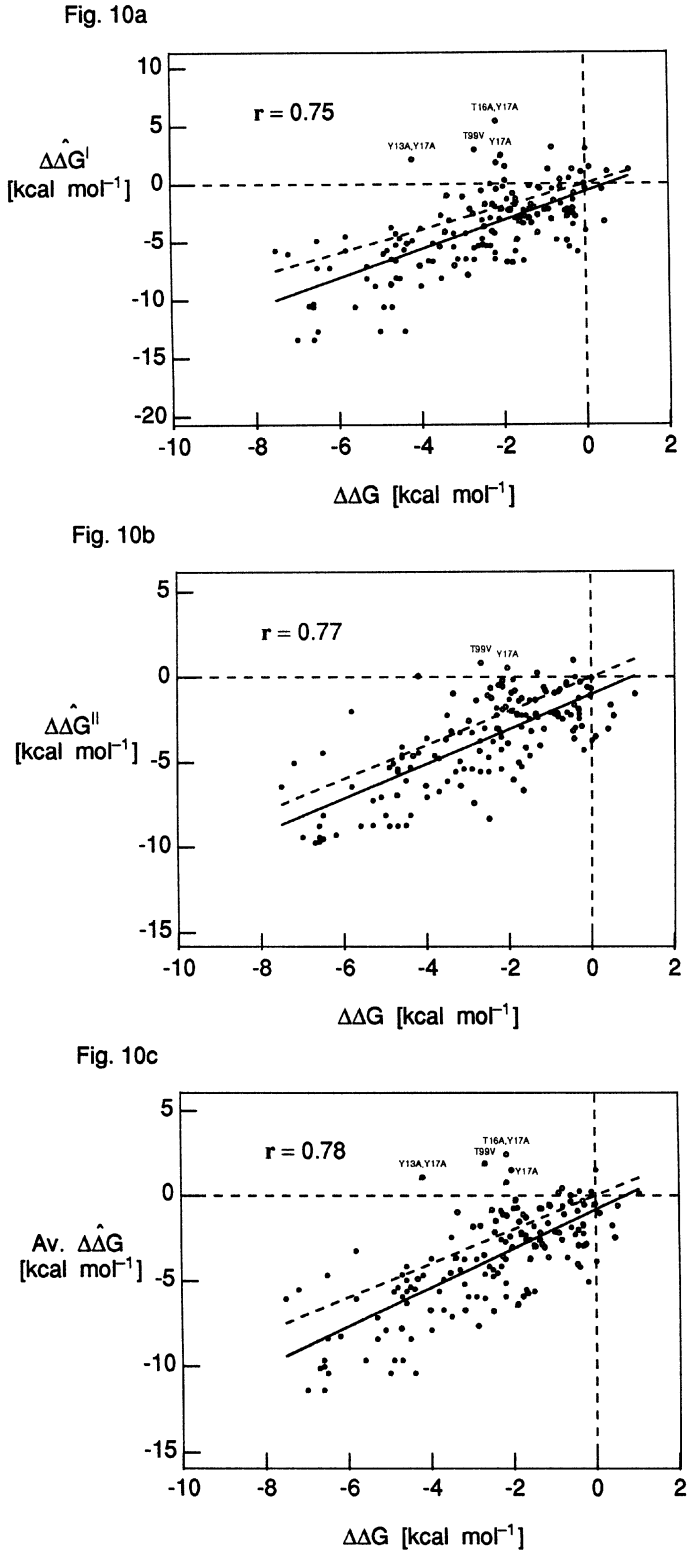

Fig. 10. Predicted stabilities ($\Delta\hat{\Delta}G$) versus experimental values of $\Delta\Delta G$ for staphylococcal nuclease and barnase modelled mutants. Calculations of (**a**) $\Delta\hat{\Delta}G^{I}$ (Method I), (**b**) $\Delta\hat{\Delta}G^{II}$ (Method II) and (**c**) their average (Av. $\Delta\Delta G$) were based on mutant residue environment determination from model atomic coordinate sets (see Materials and methods). The 83 staphylococcal nuclease (●) and 65 barnase (○) mutations correspond to those plotted in Figure 8. Estimates of $\Delta\hat{\Delta}G^{I}$ and $\Delta\hat{\Delta}G^{II}$ were obtained by linear transformation of $\Delta\Delta s^{I}$ and $\Delta\Delta s^{II}$, respectively, using the best-fit slope and intercept estimates determined for the T4 lysozyme mutants based on residue environment extrapolation (see the legend to Figure 6). Robust weighted regression least-squares fitting of the data gave correlation coefficients (r) of (a) 0.75, (b) 0.77 and (c) 0.78. The respective unweighted linear regression correlation coefficients are 0.86, 0.90 and 0.89. The best-fit regression (solid) line parameters are: (a) slope = 1.25 ± 0.09, ordinate intercept = –0.70 ± 0.29 kcal/mol; (b) slope = 1.03 ± 0.07, ordinate intercept = –1.00 ± 0.22 kcal/mol; (c) slope = 1.15 ± 0.08, ordinate intercept = –0.83 ± 0.24 kcal/mol. The idealized dashed lines have unit slope and pass through the origin.

Availability of the program

The program SDM which performs the stability predictions described in this paper is available to the academic community free of charge. The user is first requested to contact one of us (T.L.B.) to obtain a licence agreement before installing the program available in the anonymous ftp site, ftp.cryst.bbk.ac.uk. Supplementary tables of estimated stability difference scores and observed free energy differences for all of the T4 lysozyme, barnase and staphylococcal nuclease mutants are also available at the anonymous ftp site.

Acknowledgements

We thank Dr John Overington for providing the database of aligned protein structures, Prof. Peter Privalov for helpful suggestions, and Prof. Brian Matthews and Dr Walter Baase for helpful suggestions and making available preprints of their work. C.M.T. was supported by Nederlandse Unilever and N.S. was supported by Tripos Inc.

References

Alber,T. (1989) *Annu. Rev. Biochem.*, **58**, 765–798.
Alber,T., Dao-pin,S., Wilson,K., Wozniak,J.A., Cook,S.P. and Matthews,B.W. (1987) *Nature*, **330**, 41–46.
Alber,T., Bell,J.A., Dao-Pin,S., Nicholson,H., Wozniak,J.A., Cook,S. and Matthews,B.W. (1988) *Science*, **239**, 631–635.
Anderson,D.E., Hurley,J.H., Nicholson,H., Baase,W.A. and Matthews,B.W. (1993) *Protein Sci.*, **2**, 1285–1290.
Baldwin,E.P., Hajiseyedjavadi,O., Baase,W.A. and Matthews,B.W. (1993) *Science*, **262**, 1715–1718.
Bash,P.A., Singh,U.C., Langridge,R. and Kollman,P.A. (1987) *Science*, **236**, 564–568.
Becktel,W.J. and Schellman,J.A. (1987) *Biopolymers*, **26**, 1859–1877.
Bell,J.A., Becktel,W.J., Sauer,U., Baase,W.A. and Matthews,B.W. (1992) *Biochemistry*, **31**, 3590–3596.
Bernstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyer,E.F., Brice,M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) *J. Mol. Biol.*, **112**, 535–542.
Blaber,M., Lindstrom,J.D., Gassner,N., Xu,J., Heinz,D.W. and Matthews,B.W. (1993) *Biochemistry*, **32**, 11363–11373.
Blaber,M., Zhang,X.-J., Lindstrom,J.D., Pepiot,S.D., Baase,W.A. and Matthews,B.W. (1994) *J. Mol. Biol.*, **235**, 600–624.
Blaber,M., Baase,W.A., Gassner,N. and Matthews,B.W. (1995) *J. Mol. Biol.*, **246**, 317–330.
Bordo,D. and Argos,P. (1990) *J. Mol. Biol.*, **211**, 975–988.
Bordo,D. and Argos,P. (1991) *J. Mol. Biol.*, **217**, 721–729.
Bowie,J.U., Lüthy,R. and Eisenberg,D. (1991) *Science*, **253**, 164–170.
Chen,B.-L., Baase,W.A., Nicholson,H. and Schellman,J.A. (1992) *Biochemistry*, **31**, 1464–1476.
Dang,L.X., Merz,K. and Kollman,P.A. (1989) *J. Am. Chem. Soc.*, **111**, 8505–8508.
Dao-pin,S., Baase,W.A. and Matthews,B.W. (1990) *Proteins: Struct. Funct. Genet.*, **7**, 198–204.
Dao-pin,S., Sauer,U., Nicholson,H. and Matthews,B.W. (1991a) *Biochemistry*, **30**, 7142–7153.
Dao-pin,S., Alber,T., Baase,W.A., Wozniak,J.A. and Matthews,B.W. (1991b) *J. Mol. Biol.*, **221**, 647–667.
Dao-pin,S., Söderlind,E., Baase,W.A., Wozniak,J.A., Sauer,U. and Matthews,B.W. (1991c) *J. Mol. Biol.*, **221**, 873–887.
Dao-pin,S., Anderson,D.E., Baase,W.A., Dahlquist,F.W. and Matthews,B.W. (1991d) *Biochemistry*, **30**, 11521–11529.
De Fillipis,V., Sander,C. and Vriend,G. (1994) *Protein Engng*, **7**, 1203–1208.
Dill,K.A. and Shortle,D. (1991) *Annu. Rev. Biochem.*, **60**, 795–825.
Dixon,M.M., Nicholson,H., Shewchuk,L., Baase,W.A. and Matthews,B.W. (1992) *J. Mol. Biol.*, **227**, 917–933.
Eriksson,A.E., Baase,W.A., Zhang,X.J., Heinz,D.W., Blaber,M., Baldwin,E.P. and Matthews,B.W. (1992a) *Science*, **255**, 178–183.
Eriksson,A.E., Baase,W.A., Wozniak,J.A. and Matthews,B.W. (1992b) *Nature*, **355**, 371–373.
Eriksson,A.E., Baase,W.A. and Matthews,B.W. (1993) *J. Mol. Biol.*, **229**, 747–769.
Evans,P.A., Topping,K.D., Woolfson,D.N. and Dobson,C.M. (1991) *Proteins: Struct. Funct. Genet.*, **9**, 248–266.
Evans,S.V. (1993) *J. Mol. Graphics*, **11**, 134–138.
Faber,H.R. and Matthews,B.W. (1990) *Nature*, **348**, 263–266.
Fersht,A. and Winter,G. (1992) *Trends Biochem. Sci.*, **17**, 292–294.
Gilis,D. and Rooman,M. (1996) *J. Mol. Biol.*, **257**, 1112–1126.
Goa,J., Kuezera,K., Tidor,B. and Karplus,M. (1989) *Science*, **244**, 1069–1072.
Gray,T.M. and Matthews,B.W. (1987) *J. Biol. Chem.*, **262**, 16858–16864.
Grütter,M.G., Gray,T.M., Weaver,L.H., Alber,T., Wilson,K. and Matthews,B.W. (1987) *J. Mol. Biol.*, **197**, 315–329.
Guillet,V., Lapthorn,A., Hartley,R.W. and Mauguen,Y. (1993) *Structure*, **1**, 165–176.
Heinz,D.W., Baase,W.A. and Matthews,B.W. (1992) *Proc. Natl Acad. Sci. USA*, **89**, 3751–3755.
Hurley,J.H., Baase,W.A. and Matthews,B.W. (1992) *J. Mol. Biol.*, **224**, 1143–1159.
Hynes,T.R. and Fox,R.O. (1991) *Proteins: Struct. Funct. Genet.*, **10**, 92–105.
Johnson,M.S., Overington,J.P. and Blundell,T.L. (1993) *J. Mol. Biol.*, **231**, 735–752.
Jones,D.T. (1994) *Protein Sci.*, **3**, 567–574.
Jones,D.T., Taylor,W.R. and Thornton,J.M. (1992) *Nature*, **358**, 86–89.
Lee,C. (1994) *J. Mol. Biol.*, **236**, 918–939.
Lee,C. and Levitt,M. (1991) *Nature*, **352**, 448–451.
Matoushek,A., Serrano,L. and Fersht,A.R. (1994) In Pain,R.H. (ed.), *Mechanisms of Protein Folding*. IRL Press, Oxford, UK, pp. 137–159.
Matsumura,M., Becktel,W.J. and Matthews,B.W. (1988) *Nature*, **334**, 406–410.
Matsumura,M., Signor,G. and Matthews,B.W. (1989) *Nature*, **342**, 291–293.
Matthews,B.W. (1993) *Annu. Rev. Biochem.*, **62**, 139–160.
Matthews,B.W., Nicholson,H. and Becktel,W.J. (1987) *Proc. Natl Acad. Sci. USA*, **84**, 6663–6667.
Miyazawa,S. and Jernigan,R.L. (1994) *Protein Engng*, **7**, 1209–1220.
Mosteller,F. and Tukey,J.W. (1977) *Data Analysis and Regression*. Addison-Wesley, Reading, MA.
Nicholson,H., Becktel,W.J. and Matthews,B.W. (1988) *Nature*, **336**, 651–656.
Nicholson,H., Söderlind,E., Tronrud,D.E. and Matthews,B.W. (1989) *J. Mol. Biol.*, **210**, 181–193.
Nicholson,H., Anderson,D.E., Dao-pin,S. and Matthews,B.W (1991) *Biochemistry*, **30**, 9816–9828.
Nicholson,H., Tronrud,D.E., Becktel,W.J. and Matthews,B.W. (1992) *Biopolymers*, **32**, 1431–1432.
Ota,M., Kanaya,S. and Nishikawa,K. (1995) *J. Mol. Biol.*, **248**, 733–738.
Overington,J., Johnson,M.S., Šali,A. and Blundell,T.L. (1990) *Proc. R. Soc. London*, **B241**, 132–145.
Overington,J., Donnelly,D., Johnson,M.S., Šali,A. and Blundell,T.L. (1992) *Protein Sci.*, **1**, 216–226.
Pace,C.N., Laurents,D.V. and Erickson,R.E. (1992) *Biochemistry*, **31**, 2728–2734.
Pjura,P., McIntosh,L.P., Wozniak,J.A. and Matthews,B.W. (1993a) *Proteins: Struct. Funct. Genet.*, **15**, 401–415.
Pjura,P., Matsumura,M., Baase,W.A. and Matthews,B.W. (1993b) *Protein Sci.*, **2**, 2217–2225.
Prévost,M., Wodak,S.J., Tidor,B. and Karplus,M. (1991) *Proc. Natl Acad. Sci. USA*, **88**, 10880–10884.
Privalov,P.L., Tiktopulo,E.I., Yenyaminov,S.Y., Griko,Y.V., Makhatadze,G.I. and Khechinashvili,N.N. (1989) *J. Mol. Biol.*, **205**, 737–750.
Rooman,M.J. and Wodak,S.J. (1995) *Protein Engng*, **8**, 849–858.
Šali,A. and Blundell,T.L. (1990) *J. Mol. Biol.*, **212**, 403–428.
Šali,A. and Overington,J.P. (1994) *Protein Sci.*, **3**, 1582–1596.
Sauer,U.H., Dao-pin,S. and Matthews,B.W. (1992) *J. Biol. Chem.*, **267**, 2393–2399.
Serrano,L., Horovitz,A., Avron,B., Bycroft,M. and Fersht,A.R. (1990) *Biochemistry*, **29**, 9343–9352.
Serrano,L., Kellis,J.T.,Jr, Cann,P., Matouschek,A. and Fersht,A.R. (1992) *J. Mol. Biol.*, **224**, 783–804.
Shi,Y.-Y., Mark,A.E., Wang,C.-X., Hang,F., Berendsen,H.J.C. and van Gunsteren,W.F. (1993) *Protein Engng*, **6**, 289–295.
Shortle,D., Stites,W.E. and Meeker,A.K. (1990) *Biochemistry*, **29**, 8033–8041.
Sippl,M. (1990) *J. Mol. Biol.*, **213**, 859–883.
Sippl,M. (1993) *J. Comput.-Aided Mol. Design*, **7**, 473–501.
Sippl,M. (1995) *Curr. Opin. Struct. Biol.*, **5**, 229–235.
Sneddon,S.F. and Tobias,D.J. (1992) *Biochemistry*, **31**, 2842–2846.
Snedecor,G.W. and Cochran,W.G. (1980) *Statistical Methods*. 7th edition, Iowa State University Press, IA, pp. 165–166.
Sprent,P. (1989) *Applied Nonparametric Statistical Methods*, Chapman and Hall, London, UK, pp. 148–151.
Straatsma,T.P. and McCammon,J.A. (1992) *Annu. Rev. Phys. Chem.*, **43**, 407–435.
Tidor,B. and Karplus,M. (1991) *Biochemistry*, **30**, 3217–3228.
Topham,C.M., McLeod,A., Eisenmenger,F., Overington,J.P., Johnson,M.S. and Blundell,T.L. (1993) *J. Mol. Biol.*, **229**, 194–220.
Topham,C.M., Srinivasan,N., Thorpe,C.J., Overington,J.P. and Kalsheker,N.A. (1994) *Protein Engng*, **7**, 869–894.
van Gunsteren,W.F. and Mark,A.E. (1992) *J. Mol. Biol.*, **227**, 389–395.
Wodak,S.J. and Rooman,M.J. (1993) *Curr. Opin. Struct. Biol.*, **3**, 247–259.
Wong,F.C. and McCammon,J.A. (1987) In Enrehberg,A., Rigler,R., Graslund,A. and Nilsson,L. (eds), *Structure, Dynamics and Function of Biomolecules.* Spinger-Verlag, Berlin, German, pp. 51–55.
Yang,A.S. and Honig,B. (1993) *J. Mol. Biol.*, **231**, 459–474.
Zhang,X.J., Baase,W.A. and Matthews,B.W. (1991) *Biochemistry*, **30**, 2012–2017.
Zhang,X.J., Baase,W.A. and Matthews,B.W. (1992) *Protein Sci.*, **1**, 761–776.

Received April 10, 1996; revised August 21, 1996; accepted August 23, 1996