JMB

Transproteomic Evidence of a Loop-Deletion Mechanism for Enhancing Protein Thermostability

Michael J. Thompson and David Eisenberg*

UCLA-DOE Laboratory of Structural Biology and Molecular Medicine
Box 951570, University of California Los Angeles, Los Angeles, CA 90095-1570, USA

**Corresponding author*

Understanding the molecular determinants of protein thermostability is of theoretical and practical importance. While numerous determinants have been suggested, no molecular feature has been judged of paramount importance, with the possible exception of ion-pair networks. The difficulty in identifying the main determinants may have been the limited structural information available on the thermostable proteins. Recently the complete genomes for mesophilic, thermophilic and hyperthermophilic organisms have been sequenced, vastly improving the potential for uncovering general trends in sequence and structure evolution related to thermostability and, thus, for isolating the more important determinants. From a comparative analysis of 20 complete genomes, we find a trend towards shortened thermophilic proteins relative to their mesophilic homologs. Moreover, sequence alignments to proteins of known structure indicate that thermophilic sequences are more likely than their mesophilic homologs to have deletions in exposed loop regions. The new genomes offer enough comparable sequences to compute meaningful statistics that point to loop deletion as a general evolutionary strategy for increasing thermostability.

Keywords: thermophiles; genomics; proteome; surface loops

Introduction

Although the molecular basis of protein thermostability has been an area of active research for at least two decades (Perutz & Raidt, 1975; Argos *et al*., 1979), a fundamental understanding of the phenomenon remains elusive. The most obvious conclusion that can be drawn from the literature is that different proteins have adapted to different thermal environments by a variety of evolutionary devices. The lack of a fuller understanding has been due to limited data. Experimental studies comparing the structures of homologous proteins from psychrophilic (cold-adapted), mesophilic and thermophilic organisms have necessarily focused on one or a few proteins. As these studies have been relatively few in number, subsequent theoretical surveys have also been limited. Beyond the scarcity of data, our comprehension of thermostability is hampered by the large number and complexity of possible contributing factors (for recent discussions and reviews, see Vogt *et al*., 1997a,b; Ladenstein & Antranikian, 1998; Jaenicke & Böhm, 1998).

The problem with scarcity of sequences has been overcome with the complete sequencing of genomes from 20 organisms, including mesophiles, thermophiles, and hyperthermophiles. With this vast amount of data, the various suggested mechanisms for altering thermostability can be examined to see if they are, in fact, general evolutionary strategies. Although there is high-resolution structural data for only a miniscule fraction of these proteins, this limitation can be overcome somewhat by aligning a subset of the translated open reading frames (ORFs) of these 20 proteomes to proteins with known structures. With these

Abbreviations used: ORF, translated open reading frame; AA, *Aquifex aeolicus*; AF, *Archaeaoglobus fulgidis*; BB, *Borellia burgdorferi*; BS, *Bacillus subtilis*; CE, *Caenorhabditis elegans*; CT, *Chlamydia trachomatis*; EC, *Escherichia coli*; HI, *Haemophilus influenzae*; HP, *Helicobacter pylori*; MG, *Mycoplasma genitalium*; MJ, *Methanococcus jannaschii*; MP, *Mycoplasma pneumonia*; MT, *Methanobacterium thermoautotrophicum*; PA, *Pyrobaculum aerophilum*; PH, *Pyrococcus horikoshii*; RP, *Rickettsia prowazekii*; SC, *Saccharomyces cerevisiae*; SP, *Synechocystis sp.* PC6803; TB, *Mycobacterium tuberculosis*; TP, *Treponema pallidum*; SD_{54}, database of 54 structure proteins and their ORF homologs; $HOPS_{20}$, database of homologous ORF pairs from 20 complete proteomes.

E-mail address of the corresponding author: david@mbi.ucla.edu

0022-2836/99/270595–10 $30.00/0

''inferred'' structures, the structural context of observed sequence differences can be interpreted.

One possibility for increasing the thermostability of a protein is to lower its unfolding entropy, ΔS_u. This possibility has been investigated by a number of researchers using site-directed mutagenesis (Hecht *et al*., 1986; Matthews *et al*., 1987; Nicholson *et al*., 1992; Hardy *et al*., 1993; Zhang *et al*., 1995; Kawamura *et al*., 1996; Bogin *et al*., 1998; Van den Burg *et al*., 1998). Some examples of this are the substitution of other residues with the conformationally constrained proline or the replacement of glycine with less flexible residues. In addition to mutating one residue to another, the conformational entropy of the protein can be lowered by shortening the polypeptide chain.

There is evidence, both theoretical and experimental that deletion of exposed loop regions of protein structure can enhance stability. Simulations of protein unfolding have shown that unfolding beings in exposed loop regions (Daggett & Levitt, 1992; Lazaridis *et al*., 1997). Loop truncation has been a factor noted in several studies comparing crystal structures from mesophilic and thermophilic sources (Russell *et al*., 1994, 1997a; Sakon *et al*., 1996; Macedo-Ribeiro *et al*., 1996; Auerbach *et al*., 1997; Villbrandt *et al*., 1997; Tahirov *et al*., 1998). In one notable example, Usher *et al*. compared the structures of the CheY protein from *Thermotoga maritima* and *Escherichia coli*, and found no increase in ion-pairs, ion-pair networks or hydrogen-bonding (Usher *et al*., 1998). Rather, they observed a shortening of the N and C termini of the thermostable protein, truncation of one of its loops and an increase in proline residues, all entropic factors. On the low end of the temperature scale, the inverse effect has been observed. From comparisons of structures from mesophilic or thermophilic *versus* psychrophilic (cold-adapted) organisms the psychrophilic proteins are found to have insertions in exposed loop regions (Davail *et al*., 1994; Narinx *et al*., 1997; Russell *et al*., 1997b). The most compelling example of this stability-enhancing strategy is found in the work of Nagi & Regan (1997) who found a near perfect inverse correlation between loop length and stability based on designed versions of the protein Rop. Of course, there can be limitations to such a strategy. As pointed out by Robinson & Sauer (1998) in their study of linker length effects on stability of single-chain Arc repressor, truncations might interfere with biological activity or create destabilizing strain in the protein structure.

Prompted by the theoretical, experimental, and observational evidence and the wealth of proteome data at hand, we have investigated whether nature has employed loop deletion as a means of improving protein thermostability. We find that there is a statistically significant trend for sequence truncation with elevated environmental temperature. Moreover, using alignments of proteomic sequences to proteins of known structure we find that thermophilic sequences have an increased propensity for sequence deletions corresponding to surface loop regions of protein structure.

Results

Sequence lengths of thermophiles *versus* mesophiles

From all-against-all sequence alignments of the 61,947 ORFs from 20 proteomes, the average length differences for pairs of homologous sequences between each pair of proteomes was computed. These are shown as the density plot (Mathematica (Wolfram, 1996)) in Figure 1. The trend for shortness of sequences from thermophilic organisms relative to mesophilic organisms can be seen by eye (the dark blue to magenta squares in the lower right-hand rectangle).

To quantify and analyze the data presented in Figure 1, we computed the correlation coefficients between the average differences in sequence length for pairs of homologous proteins from pairs of proteomes and the differences in: (1) growth temperatures of the organisms; (2) thermophilic types of the organisms; and (3) phylogenetic classifications of the organisms. These are given in Table 1 along with the *p*-values for obtaining these correlations by chance. Notice that all three correlations are statistically significant and are consistent with the hypothesis that shorter proteins are generally more thermostable.

Principal component analysis was performed on composition vectors for all of the 20 complete proteomes to investigate the relative role of deletions compared to residue composition. The results of this analysis can be seen in Figure 2, a 2D projection of the proteome amino acid and gap composition vectors onto the two dominant axes (eigenvectors with the largest eigenvalues). The horizontal axis, which accounts for 58 % of the variance of the proteomic data, is dominated by the compositional peculiarities of some of the mesophilic proteomes (overabundance of lysine and isoleucine and depletion of alanine in the positive direction) that are not of interest to this thermostability study. However, the horizontal axis, which accounts for an additional 24 % of the total variance in the data, shows a clear separation of the mesophiles from the thermophiles. Moving down the vertical axis from mesophiles to thermophiles is associated with an increase in the fraction of deletions. In terms of amino acid composition, moving down the vertical axis is also associated with an increase in glutamate, valine, arginine and glycine and a decrease in glutamine, serine, asparagine and lysine. The implications of these relative compositional differences for thermostability is under investigation.

Given the statistically significant trend for thermophilic sequences to be shorter than their mesophilic homologs and the consequently

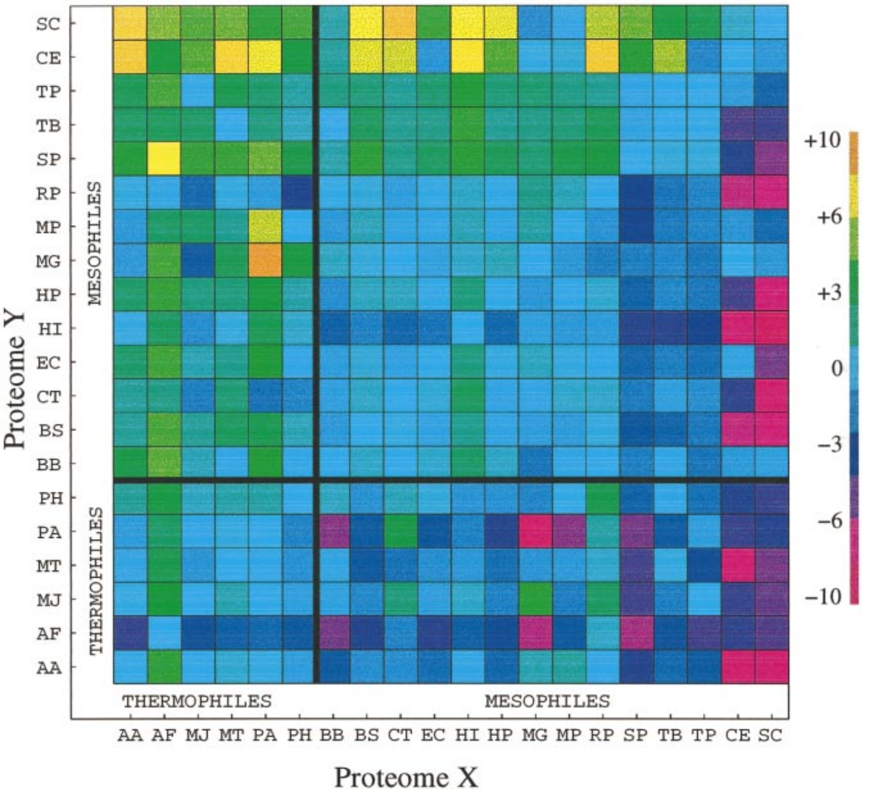

Figure 1. Sequences from thermophilic organisms tend to be shorter than their homologs from mesophilic organisms. The color of each cell (X, Y) denotes the average change in sequence length for ORFs from proteome Y relative to their homologs in proteome X. The bold lines separate the thermophiles from the mesophiles. The scale to the right shows correspondence between color and length difference in number of residues.

increased deletions, we would like to know where these deletions occur in terms of structure. This is provided by the propensities for finding deletions corresponding to each of six types of local structure as obtained from our SD_{54} dataset (see Materials and Methods). These propensities and their associated uncertainties are given in Figure 3.

Because insertions in the proteomic sequences have no aligned structures, they cannot be analyzed in the same fashion as the deletions. Unlike deletions, we find no significant differences in the fraction of insertions in mesophiles and thermophiles. The average fraction of inserted residues in mesophilic homologs and thermophilic homologs of the proteins of known structure was the same (2.4(±3)%).

Table 1. Differences in sequence length correlate more strongly with thermophily than with phylogeny

Descriptor	C	p
Δ*TEMP*	−0.52	0.3×10^{-37}
Δ*THERM*	−0.51	0.1×10^{-36}
Δ*PHYLO*	−0.48	0.2×10^{-33}

Correlation coefficients (C) between the average differences in sequence length between pairs of proteomes and the differences in their growth temperatures (Δ*TEMP*), the differences in their thermotypic classification (Δ*THERM*), or the differences in their phylogenetic classifications (Δ*PHYLO*). The probabilities of obtaining these correlations or better by change (p) are also given. See Table 2 for the values of *TEMP*, *THERM*. and *PHYLO*.

Discussion

Thermodynamics

As reviewed in the Introduction, there is theoretical, observational and experimental support that deletion of exposed loops enhances protein thermostability by lowering the entropy change of unfolding, thus raising the free energy of unfolding. In the following, we present a thermodynamic plausibility argument for this mechanism.

If protein X' has a higher thermostability than its homolog protein X, then the unfolding transition temperature of protein X', $T_u(X')$, is greater than that of protein X, $T_u(X)$:

$$T_u(X') > T_u(X) \tag{1}$$

The standard free energy difference between the denatured state and the native state at temperature, T, can be expressed as:

$$\Delta G_u^\circ = \Delta H_u - T\Delta S_u \tag{2}$$

where ΔH_u and ΔS_u are the respective changes in enthalpy and entropy for unfolding. Because native and denatured proteins are in equilibrium at T_u, the standard free energy difference is zero at this temperature. Thus, we obtain the following

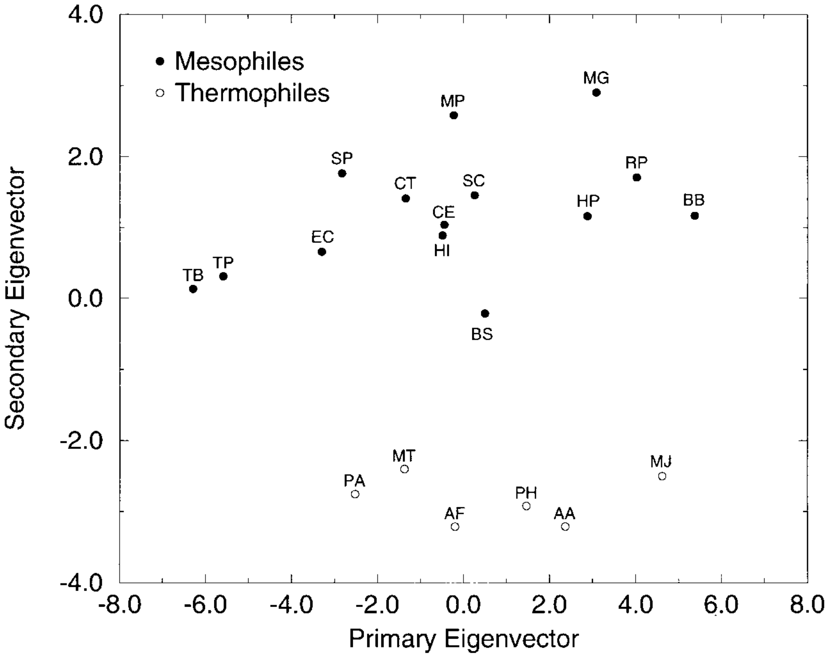

Figure 2. Thermophiles cluster based on compositional differences and deletions. Depicted in a projection of composition-difference vectors for each proteome onto the two principal axes (eigenvectors with largest eigenvalues) obtained from principal component analysis as explained in Materials and Methods. Each point is labeled by a two-letter code for the proteome, as given in Table 2. Notice that the thermophilic bacteria *Aquifex aeolicus* (AA) clusters with the archael thermophiles rather than with the mesophilic bacteria. The components (compositional changes), sorted according to magnitude, which define the vertical axis are: E (−0.53), Q (0.43), V (−0.30), S (0.28), R (−0.26), N (0.25), L (0.25), - (−0.24), T (0.17), G (−0.13), K (−0.13), H (0.12), Y (−0.11), C (0.10), F (0.09), P (−0.06), A (0.06), D (0.03), I (−0.03), W (0.02), M (−0.01).

relation at temperature T_u:

$$T_u = \frac{\Delta H_u}{\Delta S_u} \tag{3}$$

Substituting this into equation (1) yields:

$$\frac{\Delta H_u(X')}{\Delta S_u(X')} > \frac{\Delta H_u(X)}{\Delta S_u(X)} \tag{4}$$

We are interested in knowing what changes would have to be made to protein X to transform it into protein X' with higher thermostability. Thus, we are primarily interested in the differences between these unfolding enthalpy and entropy changes for the two protein forms:

$$\Delta\Delta H_u(X,\ X') = \Delta H_u(X') - \Delta H_u(X) \tag{5}$$

$$\Delta\Delta S_u(X,\ X') = \Delta S_u(X') - \Delta S_u(X) \tag{6}$$

Substituting these relationships into equation (4) we obtain:

$$\frac{\Delta H_u(X) + \Delta\Delta H_u(X,\ X')}{\Delta S_u(X) + \Delta\Delta S_u(X,\ X')} > \frac{\Delta H_u(X)}{\Delta S_u(X)} \tag{7}$$

After cross-multiplication, cancellation of terms and substitution of equation (3), we obtain the following relation:

$$\Delta\Delta H_u(X < X') > T_u(X) \times \Delta\Delta S_u(X < X') \tag{8}$$

where $T_u(X)$ is a constant. This provides a criterion for the relative changes in unfolding enthalpy and unfolding entropy necessary for increasing a protein's thermostability.

The question, then, is whether the truncation of exposed loop residues in a protein structure would fulfill the requirement expressed in

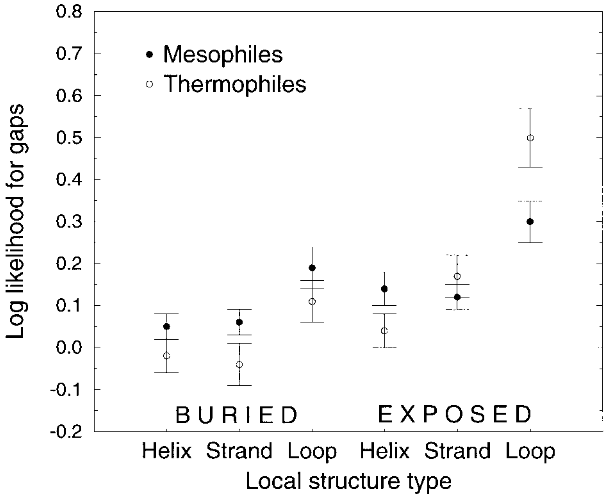

Figure 3. Thermophilic sequences have an increased propensity for deletions in exposed loops. This Figure gives the structural propensities for gaps found in alignments between proteins of known structure and their mesophilic or thermophilc homologs. Propensities less than or greater than 0 for a given structure type indicate that, respectively, fewer or more gaps are associated with that structure type than random expectation. Propensities were averaged over homologs for each protein of known structure. Standard deviations in these propensities are shown with vertical bars. Notice that exposed loops of thermophiles are the only structura elements having significantly greater gaps com-

equation (8). One component to both the ΔH_u and ΔS_u that we can neglect is that arising from the hydration of buried groups during unfolding. Because we are considering deletion of exposed residues, we expect the contributions of buried groups to be small.

We must also consider the affect of loop deletion on the contributions made by hydrogen bonds to ΔH_u (Privalov, 1979). Little change in this term would result from deletion of residues in exposed loop sites, because these structural locations imply relatively few interactions between these residues and the rest of the protein. Any interactions with the solvent would be the same in the denatured and native states, so deletion of these would have minimal impact. Barring the loss of particularly stabilizing interactions, it seems reasonable that $\Delta\Delta H_u(X, X') \approx 0$ upon removal of exposed loop residues.

Finally, we consider the contribution of conformational entropy to ΔS_u. Deletions would decrease the entropy of both the native and denatured states. However, while the decrease in conformational entropy of the native state would be small and localized to the loop itself, the deletion would have a multiplicative effect on the number of conformations for the entire chain in the unfolded state. The overall effect would be a decrease in the entropy term, $\Delta S_u(X')$, resulting in $\Delta\Delta S_u(X, X') < 0$. Given the arguments of the preceding paragraph, the condition expressed in equation (8) would be fulfilled and the thermostability of the protein would be increased.

Transproteomic evidence for sequence truncation

From our analysis of 20 complete proteomes, we have uncovered a statistically significant trend for sequence truncation in thermostable proteins compared to their mesostable homologs. This is shown in Figure 1 and Table 1. This is consistent with at least a dozen previous studies (Davail *et al.*, 1994; Sakon *et al.*, 1996; Macedo-Ribeiro *et al.*, 1996; Narinx *et al.*, 1997; Russell *et al.*, 1994, 1997a,b; Auerbach *et al.*, 1997; Villbrandt *et al.*, 1997; Wallon *et al.*, 1997; Tahirov *et al.*, 1998; Usher *et al.*, 1998) in which comparisons of individual homologous structures from psychrophilic, mesophilic, and thermophilic sources have suggested truncation or deletion of loop regions with increased temperature. It appears that these suggestions are examples of a general trend in evolutionary adaptation for high temperature environments.

Thermophily or phylogeny?

The major problem with interpreting statistical correlations between thermostability and sequence characteristics from the 20 available proteomes arises from the fact that five of the six thermophilic proteomes are from archaea and only one from a eubacterium. This leads to the question of whether the relative sequence shortening we observe in thermophiles is truly an adaptation for higher thermostability or whether it is an historical feature shared by the archaea because of their common ancestry, perhaps the result of some non-thermal adaptation.

This ambiguity can be addressed by consideration of the thermophilic bacterium, *Aquifex aeolicus*. This organism can be grouped with the thermophilic archaea (thermotypic classification) or with the mesophilic eubacteria (phylogenetic classification). If the correlation between temperature and sequence length were induced by a correlation between phylogeny and sequence length, we would expect a stronger correlation when *A. aeolicus* is grouped with its bacterial relatives. This is not the case, as seen in Table 1. Although the difference in these correlation coefficients is small, this comparison suggests that the trend for relative shortening of sequences is shared by thermophilic organisms regardless of phylogenetic lineage. It is more likely that the inverse case is true, the correlation between sequence length and phylogeny is induced by the real correlation between sequence length and thermophily through the strong overlap between the phylogenetic and thermotypic classifications.

The clustering of *A. aeolicus* with the thermophilic archaea can be seen in the principal component analysis plot of Figure 2. The Figure shows a projection of the residue composition vectors for each of the proteomes onto the two principal eigenvectors that define the residue composition space for theses 20 proteomes. One of the major components of the vertical axis that separates the organisms according to thermostability is the relative fraction of gaps (increased when moving down the axis). Sequences from thermophilic organisms show an increased fraction of gaps compared to their homologs from mesophilic organisms.

Surface loop deletion

Consistent with the shortening of the thermophilic sequences, an increased percentage of gaps in thermophilic sequences was found in all structure types (data not shown). In looking at Figure 3, we also see an increased propensity for finding gaps in exposed loop regions in thermophilic sequences. This is consistent with the thermodynamic plausibility argument presented earlier in this Discussion. Deletion of the exposed loop residues would decrease the unfolding entropy while having minimal impact on the enthalpy of unfolding.

Conclusions

We have taken a proteome-wide perspective on the problem of protein thermostability. While traditional studies have been limited to the comparison of two or a few homologs from organisms living at different temperatures, we have analyzed

45,942 homologous relationships among 19,989 ORFs from the complete set of 61,947 ORFs from the 20 proteomes. This has allowed us to detect a weak but statistically significant signal. In our analysis, we have focused on the putative stability-enhancing mechanism of loop-deletion. While this determinant has not received as much attention as ion-pair networks, we find that there is a general trend in nature for thermophilic sequences to be shorter than their mesophilic homologs. Our calculations also show that this trend is more likely due to thermostability than phylogeny. In addition, through alignments to proteins of known structure, we have found an increased occurrence of deletions in thermophilic sequences along with an increased propensity for these deletions to occur in regions of exposed loop in protein structure. Thus, it appears there exists a natural strategy for enhancing protein thermostability through truncations of exposed loop regions to lower the entropy of unfolding. This strategy could prove useful to protein engineers who have primarily relied on side-chain mutations in their efforts to improve protein thermostability.

Materials and Methods

The HOPS$_{20}$ database

The 20 proteomes used in this work are listed in Table 2. The translated open reading frames (ORFs) for each of these proteomes were obtained through links accessible through TIGR's web-site (http://www.tigr.org). All ORFs from all 20 complete proteomes were aligned against one another using a gapped version of the BLAST sequence alignment algorithm (Altschul *et al.*, 1997). As there were 61,947 ORFs in this total dataset, only alignments with E-value $\leqslant 1.6 \times 10^{-5}$ were selected for analysis.

By filtering the sequences used in the analysis, we have attempted to minimize the noise associated with other evolutionary processes. For example, we can avoid comparing sequences that have undergone radical changes in length (e.g. aligning a single-domain protein to a multi-domain protein). Therefore, we excluded pairs of homologs where either of the sequences had unaligned N or C termini with more than 15 residues. Also excluded were homolog pairs for which the difference in the number of residues internal to the aligned region was greater than 30 residues. Finally, the total length difference for the pair was restricted to 30 residues or fewer. After applying these criteria to the all-against-all proteomic ORF comparisons, we obtained a dataset of 45,942 pairwise homologous relationships containing 19,989 ORFs. A particular ORF could participate in more than one homologous pair. We denote this dataset of homologous ORF pairs from 20 proteomes as HOPS$_{20}$.

The choices for particular values of the criteria for dataset filtering given above are somewhat arbitrary. However, the results reported here were stable to the choice of these thresholds for less restrictive (larger) lengths. For more restrictive (smaller) length thresholds, the resulting lack of data yielded results with weak or no statistical significance.

Computing average length differences

The complete set of homologous ORF pairs, HOPS$_{20}$, was segregated into sets for each pair of proteomes. For example, the set (P_i, P_j) contains all the matches of ORFs from proteome P_i to their homologs in proteome P_j. For each protein in proteome P_i we computed the average difference in sequence length relative to all of its homologs in proteome P_j. Then, an average was taken over all the proteins in proteome P_i. This gave us the average

Table 2. Proteomic data used in the analysis of thermostability

ID	PROTEOME	TEMP	THERM	PHYLO	ORFs
AA	*Aquifex aeolicus* (Deckert *et al.*, 1998)	80	1	0	1522
AF	*Archaeaoglobus fulgidis* (Klenk *et al.*, 1997)	83	1	1	2409
BB	*Borellia burgdorferi* (Fraser *et al.*, 1997)	37	0	0	1638
BS	*Bacillus subtilis* (Kunst *et al.*, 1997)	30	0	0	4100
CE	*Caenorhabditis elegans* (CESC, 1998)	25	0	0	19,099
CT	*Chlamydia Trachomatis* (Stephens *et al.*, 1998)	37	0	0	894
EC	*Escherichia coli* (Blattner *et al.*, 1997)	37	0	0	4290
HI	*Haemophilus influenzae* (Fleischmann *et al.*, 1995)	37	0	0	1707
HP	*Helicobacter pylori* (Tomb *et al.*, 1997)	37	0	0	1577
MG	*Mycoplasma genitalium* (Fraser *et al.*, 1995)	37	0	0	479
MJ	*Methanococcus jannaschii* (Bult *et al.*, 1996)	83	1	1	1771
MP	*Mycoplasma pneumoniae* (Himmelreich *et al.*, 1996)	37	0	0	672
MT	*Methanobacterium thermoautotrophicum* (Smith *et al.*, 1997)	65	1	1	1871
PA	*Pyrobaculum aerophilum* (Fitz-Gibbon *et al.*, personal communication)	98	1	1	2681
PH	*Pyroccoccus horikoshii* (Kawarabayasi *et al.*, 1998)	98	1	1	2061
RP	*Rickettsia prowazekii* (Andersson *et al.*, 1998)	37	0	0	837
SC	*Saccharomyces cerevisiae* (Goffeau *et al.*, 1996)	25	0	0	6217
SP	*Synechocystis sp. PCC6803* (Kaneko *et al.*, 1996)	25	0	0	3168
TB	*Mycobacterium tuberculosis* (Cole *et al.*, 1998)	37	0	0	3924
TP	*Treponema pallidum* (Fraser *et al.*, 1998)	37	0	0	1030

ID is the two-letter identification abbreviation for the proteome of the organism in the NAME colume. TEMP is the growth temperature (°C) of the organism as found at the German Collection of Microorganisms and Cell Cultures or from TIGR (http://www.tigr.org, http://www.dsmz.de). THERM gives the thermotypic classification of the organisms (mesophillic = 0, thermophilic = 1) PHYLO gives the phylogenetic classification (non-archael = 0, archael = 1) of the organism. ORFs gives the number of open reading frames found in the publicly available lists.

difference in length for sequences in proteome P_i relative to proteome P_j. This quantity is denoted as $\langle \Delta L \rangle_{i,j}$.

Measuring correlations

The correlation between any two random variables X and Y can be quantified by the well known correlation coefficient:

$$C(X,\ Y) = \frac{\sum_i (x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\sum_i (x_i - \bar{X})^2}\sqrt{\sum_i (y_i - \bar{Y})^2}} \tag{9}$$

where $\overline{X}$ and $\overline{Y}$ are the mean values of X and Y.

The $\langle \Delta L \rangle_{i,j}$ values were computed for all pairs of (P_i, P_j) from the 20 complete proteomes, and the correlation between these average differences in protein lengths between pairs of proteomes and the respective differences in the organismal growth temperatures, $\Delta TEMP_{i,j}$ was calculated.

There is some uncertainty about what temperatures to use in this analysis, because organisms can survive over a range of temperatures. However, the organisms can be loosely clustered into mesophilic and thermophilic temperature domains (thermotypes) and labeled with the binary classification denoted by the *THERM* variable in Table 2. Therefore, we also measured the correlation between the average differences in protein lengths between pairs of proteomes and the differences, $\Delta THERM_{i,j}$, between the thermotypes of the organisms.

Moreover, since all but one of the thermophiles are archaea, apparent correlation between thermophily and sequence length might be induced by a correlation between phylogeny (given by the *PHYLO* variable in Table 2) and sequence length. To examine this possibility, the correlation between average protein length differences between pairs of proteomes and differences in phylogenetic classification, $\Delta PHYLO_{i,j}$, was computed.

To assess the significance of the measured correlations, we randomly scrambled the set of $\langle \Delta L \rangle_{i,j}$ values 10,000 times and recomputed the three correlation coefficients each time. The results of these randomizations were histogrammed and fit to Gaussian distributions. From these distributions, we calculated the probabilities of obtaining the original correlation coefficients by chance (*p*-value).

Principal component analysis

To gain some insight into the importance of deletions relative to compositional differences between thermophiles and mesophiles we performed principal component analysis on our dataset of transproteome alignments. A relative composition vector was computed for each proteome by taking the alignments of ORFs from that proteome to their homologs in the mesophilic proteomes and computing the average changes in composition for each amino acid type and the average change in fraction of gaps found in the alignments. A covariance matrix was constructed from these vectors and the eigenvectors and associated eigenvalues were calculated. The eigenvectors of the covariance matrix define the directions in this compositional space along which variance in the data is found. By sorting the eigenvectors according to the magnitude of their eigenvalues, we can determine the directions in composition space along which the most variance is found.

Inferred structural data and the SD_{54} dataset

To investigate the structural implications of the sequence changes observed between mesophiles and thermophiles, we constructed a dataset of alignments between proteomic ORFs and proteins of known structure. All ORFs from all 20 proteomes were aligned using a gapped version of BLAST to 890 chains from the PDBselect set of non-homologous representative proteins (Hobohm & Sander, 1994). Because these proteins are non-homologous, derived statistics should be free of biases toward particular protein families. The same *E*-value cut-offs and length criteria that were applied to the $HOPS_{20}$ dataset were applied to the alignments of the ORFs to these proteins of known structure. In addition, only those proteins of known structure that have both mesophilic and thermophilic homologs were included. This resulted in a dataset (SD_{54}) of 54 proteins of known structure with alignments to multiple proteomic ORFs of both thermotypes. This set contains a total of 953 pairwise alignments.

The dataset SD_{54} provides information on the structural locations of changes between mesophilic and thermophilic sequences. To gain insight into the structural significance of changes in sequence (for example, the location of deletions) secondary structures and solvent accessibilities were taken from the DSSP files for the proteins of known structure (Kabsch & Sander, 1983). Three types of secondary structure (helix, strand and loop) were considered. All sites that were not canonical helix or strand were considered as loop. Positions in the proteins of known structure were defined as buried if solvent accessibility is below 0.20, and exposed if above. Thus, there are six types of combined secondary structure and solvent accessibility categories. Residue sites in proteomic ORFs were assigned the same structure as their aligned site in the protein of known structure.

Structural propensities

From the dataset SD_{54} we are able to estimate the propensities for deletions to occur in each of the six types of local structure. For each alignment between a protein of known structure, *i*, and one of its homologs, we computed the odds ratio, $R(\phi$, gap) for gaps aligned to each of the six structure types, indexed by ϕ. Thus:

$$R_{i_{t_j}}(\phi,\ \text{gap}) = \frac{P_{i_{t_j}}(\phi,\ \text{gap})}{P_{i_{t_j}}(\phi) \times P_{i_{t_j}}(\text{gap})} \tag{10}$$

where $P(\phi$, gap) is the probability of seeing a gap aligned to structure type ϕ and $P(\phi)$ and P(gap) are the respective probabilities for observing structure type ϕ and gap at any alignment position; *t* denotes the thermotype of the homolog; and *j* is an index over the homologs of a given type *t* for protein *i*.

For each protein of known structure, we then computed two averages over $R_{i_{tj}}(\phi$, gap), one for the thermophilic homologs and one for the mesophilic homologs. Finally structural propensities for each thermotype, $S_t(\phi$, gap), were calculated by averaging over the 54 proteins of known structure:

$$S_t(\phi,\ \text{gap}) = \langle \langle R_{i_{t_j}}(\phi,\ \text{gap}) \rangle_j \rangle_i \tag{11}$$

Uncertainties for these average values were computed based on the variance observed in the average values for each protein of known structure.

Acknowledgments

The authors thank Sorel Fitz-Gibbon and Jeffrey Miller for providing sequence data for the genome of *Pyrobaculum aerophilum*. Helpful suggestions were made by Robert Scott, Todd O. Yeates, Robert Grothe, Matteo Pellegrini and Edward Marcotte. Duilio Cascio and Marek Grzeskowiak provided computer assistance. We thank DOE and NIH for support.

References

Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. (1997). Gapped BLAST and Psi-BLAST: a new generation of protein database. *Nucl. Acids Res.* **25**, 3389-3402.

Andersson, S., Zomorodipour, A., Andersson, J., Sicheritz-Ponten, T., Alsmark, C., Podowski, R., Naslund, A., Eriksson, A.-C., Winkler, H. & Kurland, C. G. (1998). The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature,* **396**, 133-140.

Argos, P., Rossman, M., Grau, U., Zuber, H., Frank, G. & Tratschin, J. (1979). Thermal stability and protein structure. *Biochemistry,* **18**, 5698-5703.

Auerbach, G., Huber, R., Grättinger, M., Zaiss, K., Schurig, H., Jaenicke, R. & Jacob, U. (1997). Closed structure of phosphoglycerate kinase from *Thermotoga maritima* reveals the catalytic mechanism and determinants of thermal stability. *Structure,* **5**, 1475-1483.

Blattner, F., Plunkett, G., Bloch, C., Perna, N., Burland, V., Riley, M., Collado-Vides, J., Glasner, J., Rode, C., Mayhew, G., Gregor, J., Davis, N., Kirkpatrick, H., Goeden, M. & Rose, D., *et al.* (1997). The complete genome sequence of *Escherichia coli* K-12. *Science,* **277**, 1453-1474.

Bogin, O., Peretz, M., Hacham, Y., Korkhin, Y., Frolow, F., Kalb, (Gilboa) A. & Burstein, Y. (1998). Enhanced thermal stability of *Clostridium beijerinckii* alcohol dehydrogenase after strategic substitution of amino acid residues with prolines from the homologous thermophilic *Thermoanaerobacter brokii* alcohol dehydrogenase. *Protein Sci.* **7**, 1156-1163.

Bult, C., White, O., Olsen, G., Zhou, L., Fleischmann, R., Sutton, G., Blake, J., FitzGerald, L., Clayton, R., Gocayne, J., Kerlavage, A., Dougherty, B., Tomb, J., Adams, M. & Reich, C., *et al.* (1996). Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science,* **273**, 1058-1073.

CESC (The *Caenorhabditis elegans* Sequencing Consortium) (1998). Genome sequence of the nematode *Caenorhabditis elegans*: a platform for investigating biology. *Science,* **282**, 2012-2018.

Cole, S., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S., Eiglmeier, K., Gas, S., Barry, C., Tekaia, F., Badcock, K., Basham, D., Brown, D. & Chillingworth *et al.*, (1998). Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature,* **393**, 537-544.

Daggett, & Levitt, M. (1992). A model of the molten globule state from molecular dynamics simulations. *Proc. Natl Acad. Sci. USA,* **89**, 5142-5146.

Davail, S., Feller, G., Narinx, E. & Gerday, C. (1994). Cold adaptation of proteins. Purification, characterization, and sequence of the heat-labile subtilisin from the antarctic psychrophile *Bacillus* TA41. *J. Biol. Chem.* **269**, 17448-17453.

Deckert, G., Warren, P., Gaasterland, T., Young, W., Lenox, A., Graham, D., Overbeek, R., Snead, M., Keller, M., Aujay, M., Huber, R., Feldman, R., Short, J., Olsen, G. & Swanson, R. (1998). The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. *Nature,* **392**, 353-358.

Fleischmann, R., Adams, M., White, O., Clayton, R., Kirkness, E., Kerlavage, A., Bult, C., Tomb, J., Dougherty, B., Merrick, J., McKenney, K., Sutton, G., FitzHugh, W., Fields, C. & Gocayne, J., *et al.* (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science,* **269**, 496-512.

Fraser, C., Gocayne, J., Whiate, O., Adams, M., Clayton, R., Fleischmann, R., Bult, C., Kerlavage, A., Sutton, G., Kelly, J., Fritchman, J., Weidman, J., Small, K., Sandusky, M. & Fuhrmann, J., *et al.* (1995). The minimal gene complement of *Mycoplasma genitalium*. *Science,* **270**, 397-403.

Fraser, C., Casjens, S., Huang, W., Sutton, G., Clayton, R., Lathigra, R., White, O., Ketchum, K., Dosons, R., Hickey, E., Gwinn, M., Dougherty, B., Tomb, J., Fleischmann, R. & Richardson, D., *et al.* (1997). Genomic sequence of a lyme disease spirochaete, *Borrelia burgdorferi*. *Nature,* **390**, 580-586.

Fraser, C., Norris, S., Weinstock, G., White, O., Sutton, G., Dodson, R., Gwinn, M., Hickey, E., Clayton, R., Ketchum, K., Sodergren, E., Hardham, J., McLeod, M., Salzberg, S. & Peterson, J. *et al.* (1998). Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. *Science,* **281**, 375-388.

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J., Jacq, C., Johnston, M., Louis, E., Mewes, H., Murakami, Y., Philippsen, P., Tettelin, H. & Oliver, S. (1996). Life with 6000 genes. *Science,* **274**, 546-567.

Hardy, F., Vriend, G., Veltman, O., van der Vinne, B., Venema, G. & Eijsink, V. (1993). Stabilization of *Bacillus stearothermophilus* neutral protease by introduction of prolines. *FEBS Letters,* **317**, 89-92.

Hecht, M., Sturtevant, J. & Sauer, R. (1986). Stabilization of lambda repressor against thermal denaturation by site-directed Gly-Ala changes in alpha helix 3. *Proteins: Struct. Funct. Genet.* **1**, 43-46.

Himmelreich, R., Hilbert, H., Plagens, H., Pirkl, E., Li, B. & Herr, R. (1996). Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucl. Acids Res.* **24**, 4420-4449.

Hobohm, U. & Sander, C. (1994). Enlarged representative set of protein structures. *Protein Sci.* **3**, 522-524.

Jaenicke, R. & Böhm, G. (1998). The stability of proteins in extreme environments. *Curr. Opin. Struct. Biol.* **8**, 738-748.

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

Kaneko, T., Sato, S., Kotani, H., Tanaka, A., Asamizu, E., Nakamura, Y., Miyajima, N., Hirosawa, M., Sugiura, M., Sasamoto, S., Kimura, T., Hosouchi, T., Matsuno, A., Muraki, A. & Nakazaki, N., *et al.*

(1996). Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis sp. strain* PCC6803. II. sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res.* **3**, 109-136.

Kawamura, S., Kakuta, Y., Tanaka, I., Hikichi, K., Kuhura, S., Yamasaki, N. & Kimura, M. (1996). Glycine-15 in the bend between two alpha-helices can explain the thermostability of DNA binding protein Hu from *Bacillus stearothermophilus*. *Biochemistry,* **35**, 1195-1200.

Kawarabayasi, Y., Sawada, M., Horikawa, H., Haikawa, Y., Hino, Y., Yamamoto, S., Sekine, M., Baba, S., Kosugi, H., Hosoyama, A., Nagai, Y., Sakai, M., Ogura, K., Otsuka, R. & Nakazawa, H., *et al*. (1998). Complete sequence and gene organization of the genome of a hyper-thermophilic archaebacterium *Pyrococcus horikishii* OT3. *DNA Res.* **5(2)**, 55-76.

Klenk, H., Clayton, R., Tomb, J., White, O., Nelson, K., Ketchum, K., Dodson, R., Gwinn, M., Hickey, E., Peterson, J., Richardson, D., Kerlavage, A., Graham, D., Kyrpides, N. & Fleischmann, R., *et al*. (1997). The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. *Nature,* **390**, 364-370.

Kunst, F., Ogasawara, N., Moszer, I., Albertini, A., Alloni, G., Azevedo, V., Bertero, M., Bessieres, P., Bolotin, A., Borchert, S., Borriss, R., Boursier, L., Brans, A., Braun, M. & Brignell, S., *et al*. (1997). The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. *Nature,* **390**, 249-256.

Ladenstein, R. & Antranikian, G. (1998). Proteins from hyperthermophiles: stability and enzyme catalysis close to the boiling point of water. *Adv. Biochem. Eng. Biotechnol.* **61**, 37-85.

Lazaridis, T., Lee, I. & Karplus, M. (1997). Dynamics and unfolding pathways of a hyperthermophilic and a mesophilic rubredoxin. *Protein Sci.* **6**, 2589-2605.

Macedo-Ribeiro, S., Darimont, B., Sterner, R. & Huber, R. (1996). Small structural changes account for the high thermostability of 1[4Fe-4S] ferredoxin from the hyerthermophilic bacterium *Thermotoga maritima*. *Structure,* **4**, 1291-1301.

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

Nagi, A. & Regan, L. (1997). An inverse correlation between loop length and stability in a four-helix bundle protein. *Folding Design,* **2**, 67-75.

Narinx, E., Baise, E. & Gerday, C. (1997). Subtilisin from psychrophilic antarctic bacteria: characterization and site-directed mutagenesis of residues possibly involved in the adaptation to cold. *Protein Eng.* **10**, 1271-1279.

Nicholson, H., Tronrud, D., Becktel, W. & Matthews, B. (1992). Analysis of the effectiveness of proline substitutions and glycine replacements in increasing the stability of phage T4 lysozyme. *Biopolymers,* **32**, 1421-1441.

Perutz, M. & Raidt, H. (1975). Stereochemical basis of heat stability in bacterial ferredoxins and in haemoglobin A2. *Nature,* **255**, 256-259.

Privalov, P. (1979). Stability of proteins: small globular proteins. *Advan. Protein Chem.* **33**, 167-241.

Robinson, C. & Sauer, R. (1998). Optimizing the stability of single-chain proteins by linker length and composition mutagenesis. *Proc. Natl Acad. Sci. USA,* **95**, 5929-5934.

Russell, R., Hough, D., Danson, M. & Taylor, G. (1994). The crystal structure of citrate synthase from the thermophilic archaeon, *Thermoplasma acidophilum*. *Structure,* **2**, 1157-1167.

Russell, R., Ferguson, J., Hough, D., Danson, M. & Taylor, G. L. (1997a). The crystal structure of citrate synthase from the hyperthermophilic archaeon *Pyrococcus furiosus* at 1.9 Å resolution. *Biochemistry,* **36**, 9983-9994.

Russell, R., Gerike, U., Danson, M., Hough, D. & Taylor, G. L. (1997b). Structural adaptations of the cold-active citrate synthase from an antarctic bacterium. *Structure,* **6**, 351-361.

Sakon, J., Adney, W., Himmel, M., Thomas, S. & Karplus, P. (1996). Crystal structure of thermostable family 5 endocellulase E1 from *Acidothermus cellulolyticus* in complex with cellotetraose. *Biochemistry,* **35**, 10648-10660.

Smith, D., Doucette-Stamm, L., Deloughery, C., Lee, H., Dubois, J., Aldredge, T., Bashirzadeh, R., Blakely, D., Cook, R., Gilbert, K., Harrison, D., Hoang, L., Keagle, P., Lumm, W., Pothier, B., Quiu, D., Spadafora, R., Vicaire, R., Wang, Y., Wierzbowski, J., Gibson, R., Jiwani, N., Caruso, A., Bush, D. & Reeve, J., *et al*. (1997). Complete genome sequence of *Methanobacterium thermoautotrophicum* ΔH: functional analysis and comparative genomics. *J. Bacteriol.* **179**, 7135-7155.

Stephens, R., Kalman, S., Lammel, C., Fan, J., Marathe, R., Aravind, L., Mitchell, W., Olinger, L., Tatusov, R., Zhao, Q., Koonin, E. & Davis, R. (1998). Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*. *Science,* **282**, 754-759.

Tahirov, T., Oki, H., Tsukihara, T., Ogasahara, K., Yutani, K., Ogata, K., Izu, Y., Tsunasawa, S. & Kato, I. (1998). Crystal structure of methionine aminopeptidase from hyperthermophile, *Pyrococcus furiosus*. *J. Mol. Biol.* **284**, 101-124.

Tomb, J., White, O., Kerlavage, A., Clayton, R., Sutton, G., Fleischmann, R., Ketchum, K., Klenk, H., Gill, s. , Dougherty, B., Nelson, K., Quackenbush, J., Zhou, L., Kirkness, E. & Peterson, S., *et al*. (1997). The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature,* **388**, 539-547.

Usher, K., De la Cruz, A., Dahlquist, F., Swanson, R., Simon, M. & Remington, S. (1998). Crystal structures of CheY from *Thermotoga maritima* do not support conventional explanations for the structural basis of enhanced stability. *Protein Sci.* **7**, 403-412.

Van den Burg, B., Vriend, G., Veltman, O., Venema, G. & Eijsink, V. (1998). Engineering an enzyme to resist boiling. *Proc. Natl Acad. Sci. USA,* **95**, 2056-2060.

Villbrandt, B., Sagner, G. & Schomburg, D. (1997). Investigations on the thermostability and function of truncated *Thermus aquaticus* DNA polymerase fragments. *Protein Eng.* **10**, 1281-1288.

Vogt, G., Woell, S. & Argos, P. (1997a). Protein thermal stability, hydrogen bonds, and ion pairs. *J. Mol. Biol.* **269**, 631-643.

Vogt, G., Woell, S. & Argos, P. (1997b). Protein thermal stability: hydrogen bonds or internal packing. *Folding Design,* **1**, S40-S46.

Wallon, G., Kryger, G., Lovett, S., Oshima, T., Ringe, D. & Petsko, G. (1997). Crystal structures of *Escherichia coli* and Salmonella typhimurium 3-isopropylmalate dehydrogenase and comparison with the thermophilic counterpart from *Thermus thermophilius*. *J. Mol. Biol.* **266**, 1016-1031.

Wolfram, S. (1996). *The Mathematica Book*, 3rd edit., Wolfram Media/Cambridge University Press, Cambridge, UK.

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

Edited by I. B. Honig

(Received 11 March 1999; received in revised form 20 May 1999; accepted 20 May 1999)