Protein Engineering vol.13 no.3 pp.179–191, 2000

Factors enhancing protein thermostability

Sandeep Kumar[1], Chung-Jung Tsai[2] and Ruth Nussinov[1,3,4]

[1]Intramural Research Support Program, SAIC Frederick, [2]Laboratory of Experimental and Computational Biology, National Cancer Institute, Frederick Cancer Research and Development Center, Bldg 469, Rm 151, Frederick, MD 21702, USA and [3]Sackler Institute of Molecular Medicine, Sackler School of Medicine, Tel Aviv University, Tel Aviv 69978, Israel

[4]To whom correspondence should be addressed
Email: ruthn@ncifcrf.gov

Several sequence and structural factors have been proposed to contribute toward greater stability of thermophilic proteins. Here we present a statistical examination of structural and sequence parameters in representatives of 18 non-redundant families of thermophilic and mesophilic proteins. Our aim was to look for systematic differences among thermophilic and mesophilic proteins across the families. We observe that both thermophilic and mesophilic proteins have similar hydrophobicities, compactness, oligomeric states, polar and non-polar contribution to surface areas, main-chain and side-chain hydrogen bonds. Insertions/deletions and proline substitutions do not show consistent trends between the thermophilic and mesophilic members of the families. On the other hand, salt bridges and side chain–side chain hydrogen bonds increase in the majority of the thermophilic proteins. Additionally, comparisons of the sequences of the thermophile–mesophile homologous protein pairs indicate that Arg and Tyr are significantly more frequent, while Cys and Ser are less frequent in thermophilic proteins. Thermophiles both have a larger fraction of their residues in the α-helical conformation, and they avoid Pro in their α-helices to a greater extent than the mesophiles. These results indicate that thermostable proteins adapt dual strategies to withstand high temperatures. Our intention has been to explore factors contributing to the stability of proteins from thermophiles with respect to the melting temperatures (T_m), the best descriptor of thermal stability. Unfortunately, T_m values are available only for a few proteins in our high resolution dataset. Currently, this limits our ability to examine correlations in a meaningful way.

Keywords: melting temperature/sequence/structure/thermophiles/thermostability

Introduction

Several organisms, mainly archaea, thrive under extreme environmental conditions, e.g. high pressure in deep sea vents, high temperature and non-physiological pH found in submarine hydrothermal areas, continental sulfataras, low temperatures in Antarctica and high salt concentration in the Dead Sea and in the Great Salt Lake, and in man made geothermal power plants. There has been a growing interest in understanding the stabilization of proteins from these organisms. Such an understanding, especially of the thermophilic proteins, is not only essential for a theoretical description of the physicochemical principles behind protein folding and stability, but is also critical for designing efficient enzymes that can work at high temperatures. Such enzymes may be useful for several industrial applications, such as detergent manufacturing, food and starch processing, production of high fructose corn syrup and PCR (Adams and Kelly, 1995). It has also been noticed that thermophilic enzymes are more resistant to proteolysis than their mesophilic homologues (Daniel *et al*., 1982), probably owing to their greater rigidity.

Thermostable proteins maintain their activities and are stable at high temperatures. Identifying and understanding the factors contributing to the stability of proteins from organisms living under extreme conditions has been a long standing problem. The first high resolution crystal structure of thermolysin was reported in 1974 (Matthews *et al*., 1974). Perutz and Raidt (1975) commented on the stereochemical basis of thermostability of ferredoxins and hemoglobin A2. Since these pioneering efforts, several investigators have focused on the problem of the molecular basis of protein thermostability. Several reasons have been attributed to the greater stability of the thermophilic proteins (Querol *et al*., 1996; Jaenicke and Bohm, 1998; Ladenstein and Antranikian, 1998). Among the most prominent ones are greater hydrophobicity (Haney *et al*., 1997), better packing, deletion or shortening of loops (Russell *et al*., 1997), smaller and less numerous cavities, increased surface area buried upon oligomerization (Salminen *et al*., 1996), amino acid substitutions within and outside the secondary structures (Zuber, 1988; Haney *et al*., 1997; Russell *et al*., 1998), increased occurrence of proline residues (Haney *et al*., 1997; Watanabe *et al*., 1997; Bogin *et al*., 1998), decreased occurrence of thermolabile residues (Russell *et al*., 1997), increased helical content, increased polar surface area (Haney *et al*., 1997; Vogt and Argos, 1997; Vogt *et al*., 1997), increased hydrogen bonding (Vogt and Argos, 1997; Vogt *et al*., 1997) and salt bridges (Yip *et al*., 1995, 1998; Haney *et al*., 1997; Russell *et al*., 1997, 1998; Elcock, 1998; Xiao and Honig, 1999; Kumar *et al*., 2000).

Here we present a statistical analysis of parameters thought to contribute toward protein thermostability. We have carried out structural comparisons to cluster the thermophile–mesophile protein families, creating a non-redundant dataset of 18 families from the Protein Data Bank (PDB) (Bernstein *et al*., 1977). These families span an entire spectrum, containing proteins from moderately thermophilic to hyperthermophilic organisms and their mesophilic homologs. Not all the differences observed between the thermophilic and mesophilic proteins are due to thermostability. Here we select one pair from each family. We choose the structurally most similar thermophile–mesophile pair having the best resolution, so that the observed differences can be expected to be mostly due to thermostability. In our dataset, no two thermophilic proteins from different families have similar three-dimensional structures, ensuring a bias free sample. Between each

Table I. Families of thermophilic and mesophilic proteins

Protein family name and stability of thermophilic protein	Thermophilic organism and T_L(°C)	PDB entry, resolution (Å), oligomeric state and N_{res}	Mesophilic organism and T_L (°C)	PDB entry, resolution (Å), oligomeric state and N_{res}	R.m.s.d. (Å)	ID
Citrate synthase[a] Half life of 170.0 min at 100°C	*Pyrococcus furiosus* 100	1AJ8 1.9 Dimer 741	Chicken Heart 37	1CSH 1.6 Dimer 870	1.68	26.2
Malate dehydrogenase[b] Fully active after 60 min at 90°C	*Thermus flavus* 70–75	1BDM 2.5 Dimer 644	Porcine 37	4MDH 2.5 Dimer 666	0.94	54.1
Rubredoxin[c] Stable for >24 h at 95°C T_m = 176–195°C	*Pyrococcus furiosus* 100	1CAA 1.8 Monomer 53	*Desulfovibrio vulgaris* 34–37	8RXN 1.0 Monomer 52	0.69	66.7
Cyclodextrin glucanotransferase (CGTase)[d] >90% catalytic activity when kept at 80°C for 5 h	*Thermoanaerobacterium thermosulfurigenes* 60	1CIU 2.3 Monomer 683	*Bacillus circulans* 30–40	1CDG 2.0 Monomer 686	0.7	70.5
EF-TU and EF-TU–TS complex[e] Temperature optimum ~70°C	*Thermus aquaticus* 70–72	1EFT 2.5 Monomer 405	*Escherichia coli* 37	1EFU C 2.5 A_2B_2 Tetramer 1290 (363 in chain C)	1.5	57.6
Glutamate dehydrogenase[f] Half life of 12 h at 100°C T_m = 113°C	*Pyrococcus furiosus* 75–100	1GTM 2.2 Hexamer 2502	*Clostridium symbiosum* 30–37	1HRD 1.96 Hexamer 2694	1.38	34.3
Lactate dehydrogenase[g] Active for 30 min at 80°C	*Bacillus stearothermophilus* 40–65	1LDN 2.5 Tetramer 1264	*Plasmodium falciparum* 37	1LDG 1.74 Tetramer 1260	1.25	28.4
Thermolysin and neutral protease[h] 50% activity after 1 h at 80°C	*Bacillus thermoproteolyticus* 52.5	1LNF 1.7 Dimer 634	*Bacillus cereus* 30	1NPC 2.0 Monomer 317	0.86	73.3
3-Phosphoglycerate kinase (PGK)[i] T_m = 67°C	*Bacillus stearothermophilus* 40–65	1PHP 1.65 Monomer 394	*Saccharomyces cerevisiae* 25–30	1QPG 2.4 Dimer 830	1.28	51.4
Dimerization domain of EF-TS and EF-TU–TS complex[j] Does not denature up to 95°C	*Thermus thermophilus* 70–75	1TFE 1.7 Dimer 284	*Escherichia coli* 37	1EFU B 2.5 A_2B_2 Tetramer 1290(282 in chain B)	1.24	40.8
CheY[k] T_m = 95°C, $\Delta H°$=78 kcal/mol. Optimum temperature = 90°C	*Thermotoga maritima*	1TMY 1.9 Monomer 118	*Escherichia coli* 37	3CHY 1.66 Monomer 128	1.39	28.6
Methionine aminopeptidase[l] Half life of 4.5 h at 90°C	*Pyrococcus furiosus* 100	1XGS 1.75 Dimer 590	*Escherichia coli* 37	1MAT 2.4 Monomer 263	1.39	30.6
Endo-1,4-b Xylanase[m] Highest activity at 65°C for 15 min reaction	*Thermomyces lanuginosus* 50	1YNA 1.55 Monomer 193	*Bacillus circulans* 30–40	1XNB 1.49 Monomer 185	1.14	50.9
Adenylate kinase[n] T_m = 74.5°C ΔH = 145 kcal/mol	*Bacillus stearothermophilus* 40–65	1ZIN 1.65 Monomer 217	*Sacchromyces cerevisae* 25–30	1AKY 1.63 Monomer 218	1.22	42.0

Table 1 continued

Table I. continued

Protein family name and stability of thermophilic protein	Thermophilic organism and T_L(°C)	PDB entry, resolution (Å), oligomeric state and N_{res}	Mesophilic organism and T_L (°C)	PDB entry, resolution (Å), oligomeric state and N_{res}	R.m.s.d. (Å)	ID
Ferredoxin[o]	*Bacillus thermoproteolyticus* 52.5	2FXB 2.3 Monomer 81	*Clostridium acidurici* 19–37	1FCA 1.8 Monomer 55	1.27	24.0
Inorganic pyrophosphatase (Hydrolase)[p] Retains 50% of initial activity after 1 h at 90°C	*Thermus thermophilus* 70–75	2PRD 2.0 Hexamer 1044	*Escherichia coli* 37	1INO 2.2 Hexamer 1050	1.10	48.5
Manganese superoxide dismutase[q]	*Thermus thermophilus* 70–75	3MDS 1.8 Tetramer 812	*Homo sapiens* 37	1QNM 2.3 Tetramer 792	1.17	53.2
Phosphofructokinase[r]	*Bacillus stearothermophilus*40–65	3PFK 2.4 Tetramer 1276	*Escherichia coli* 37	2PFK 2.4 Tetramer 1208	0.87	57.1

T_L stands for living temperature, while T_m indicates melting temperature. N_{res} gives number of residues in the whole protein. R.m.s.d. stands for root mean square deviation and ID indicates sequence identity. R.m.s.d. and ID were computed for individual chains in thermophilic and mesophilic proteins. Values represent the best matches.

[a]Best match was obtained between chains B of 1AJ8 and 1CSH (Russell *et al*., 1997).

[b]Best match was obtained between chains B of 1BDM and 4MDH (Kelly *et al*., 1993).

[c]There is more than one estimate of T_m for rubredoxin (Day *et al*., 1992). The one used here is from Hiller *et al*. (1997).

[d]Knegtel *et al*. (1996).

[e]1EFU corresponds to 1EFT and 1TFE in the thermophilic proteins. Best match for 1EFT was obtained with chain C of 1EFU (Kjeldgaard *et al*., 1993).

[f]Best match was obtained between chain B of 1GTM and chain B of 1HRD (Yip *et al*., 1995). Value of T_m for 1GTM was obtained from Klump *et al*. (1992).

[g]Crystal asymmetric unit of 1LDN contains two copies of the molecule (Wigley *et al*., 1992). The first copy was used. Best match was obtained between chain C of 1LDN and 1LDG.

[h]Best match was obtained between chain E of 1LNF and 1NPC (Matthews *et al*., 1974; Holland *et al*., 1995); activity data is from Singleton and Sainsbury (1978).

[i]T_m for mesophilic enzyme = 53°C. $\Delta\Delta G$ = ~5 kcal/mol (Davies *et al*., 1993; Auerbach *et al*., 1997).

[j]Best match for 1TFE was obtained with chain B of 1EFU (Jiang *et al*., 1996).

[k]Usher *et al*. (1998).

[l]Best match was obtained between chain B of 1XGS and 1MAT (Tsunasawa *et al*., 1997).

[m]Data on activity was taken from Gomes *et al*. (1993).

[n]T_m for mesophilic adenylate kinase is 48°C. ΔH_m = 340 kJ/mol (Glaser *et al*., 1992).

[o]Fukuyama *et al*. (1988).

[p]Best match was obtained between the chains given in the asymmetric units of 2PRD and 1INO (Obmolova *et al*., 1993; Salminen *et al*., 1996).

[q]Asymmetric unit of 1QNM contains two identical chains of 198 residues each. A match was found to be the best when both the chains of 1QNM are simultaneously aligned with the chain in the asymmetric unit of 3MDS.

[r]Rypniewski and Evans (1989).

thermophile–mesophile pair, we have compared several structural properties such as oligomeric state, insertion/deletion of residues, compactness, hydrophobicity, helical content, hydrogen bonds and salt bridges. We find that most of these do not show consistent trends across the families, indicating versatile protein stabilization strategies adopted by the individual families. However, there are a few global trends across a large number of families. Salt bridges and side-chain hydrogen bonds increase in most of the thermophilic proteins. Interestingly, the overall amino acid distributions in the thermophilic and the mesophilic proteins are significantly different, in spite of the high sequence homologies between the protein structural pairs. The proportions of the thermolabile residue Cys and of Ser decrease significantly, while those of Arg and Tyr increase significantly in the thermophilic proteins as compared with their mesophilic homologs. Pro is observed to occur less frequently in α-helices of the thermophilic proteins. On the whole, a higher proportion of amino acids in the thermophilic proteins adopt α-helical conformation. Our results indicate a two pronged strategy adopted by the thermophiles. Thermophilic proteins appear to disfavor potentially destabilizing factors along with favoring the potentially stabilizing ones. Furthermore, here we compare our results with those obtained from an analysis of a database of 165 non-homologous proteins.

Our intention was to carry out the analysis with respect to the melting temperatures of the corresponding proteins, from both the thermophiles and the mesophiles. Melting temperatures (T_m's), are the best descriptor of thermal stability. To be able to draw reliable conclusions, we wished to focus on cases where (i) high resolution crystal structures are available for both the thermophilic protein and its mesophilic homolog; and (ii) melting temperatures for the thermophilic and mesophilic proteins have been measured and reported. Cases where the difference between the melting temperatures of the thermophilic–mesophilic protein pair is not too small, and that the size of the protein is large enough, are the more meaningful ones. Too small a difference in the melting temperatures corresponds to a small difference in energy between the pair of proteins; whereas if the protein is small, the differences in structural parameters might be difficult to gauge accurately. Unfortunately, only a few cases are currently available in the

literature. In these cases, the difference in the number of salt bridges between the thermophile and its mesophile homologue appears to correlate with the T_m of the thermophilic protein. While other structural factors, such as compactness and hydrophobicity, contribute to thermostability, no consistent correlation with the T_m is observed. However, we are unable to obtain statistically reliable results due to the sparse data. On the other hand, we point out that none of the structural factors correlates with the living temperatures of the thermophilic organisms.

Materials and methods

Construction of the families of thermophilic and mesophilic proteins

An index file, called source.idx, in the Protein Data Bank (PDB) (Bernstein *et al*., 1977) contains the names of the organisms for all protein crystal structures available in the PDB. The January 7, 1998 update of this file was searched for the keywords THERM and PYRO. This search yielded 167 (out of 6751) PDB entries containing different proteins from thermophilic organisms. The entries in which protein structures had been determined by using nuclear magnetic resistance (NMR) and/or theoretical modeling, $R = -1.0$ Å in cmpd_res file, were discarded, leaving us with 145 PDB entries. From this set of entries containing proteins whose structures were determined by X-ray crystallography, 113 entries containing high resolution ($R \leqslant 2.5$ Å) structures for 55 different thermophilic proteins were selected for further study. For each of the thermophilic proteins in the list, the PDB entry with the best resolution was picked. Three-dimensional structures of the thermophilic proteins were compared all against all using a sequence order independent structural comparison technique (Tsai *et al*., 1996). This computer vision-based technique superimposes spatially equivalent regions in two proteins without regard to their sequential connectivity, or to the number of residues in the protein. Since the mesophilic and thermophilic proteins have different sizes and may have different oligomeric states, this technique allows us to superimpose the conserved regions of the proteins independently of these factors. Two proteins are considered to be dissimilar if (i) the backbone C_α atom superposition for the two structures yields an r.m.s.d. $\geqslant 2.00$ Å; and (ii) the sequence identity (ID) for the two proteins is $\leqslant 20\%$. Finally, thermophilic proteins were retained in the database if they have dissimilar structures and if there is at least one high resolution crystal structure for their corresponding mesophilic homologs. This step ensures non-redundancy in the database. Eighteen different thermophilic proteins were obtained. The structure of each of the 18 proteins was compared with their corresponding homologous PDB entries. Two structures were considered to be similar if they did not satisfy both of the above conditions. At this stage, many families contain several mesophilic proteins. Application of a 2.5 Å resolution cut-off substantially decrease their number. Finally, the PDB entry which has the best resolution and contains the structure that is most similar to the thermophilic protein is selected. As far as possible, we have tried to select wild-type thermophile–mesophile pairs. Attention was also paid to the presence (absence) of substrates in the thermophilic and mesophilic proteins. Choosing one thermophile–mesophile pair per family, in a way such that the pair contains the best resolved structures along with the largest sequence and structure homology among the various available alternates, has several advantages. First, since the two proteins are most similar, the observed differences can be correlated with thermostability with a greater degree of confidence. Second, the variability, or the consistency of the results, can be judged from the behavior of all 18 families; and third, in particular, the behavior of the parameters is a function of two factors: the extent of structural similarity between the two molecules and the sequence similarity. The non-polar buried surface area, compactness, etc. obtained in comparisons of members of the same family would need to be calibrated against the sequence differences, and it is unclear how best to do this in practice. In an extensive recent analysis, Vogt *et al*. (1997) have used multiple mesophilic homologs for comparison with the thermophilic proteins. They have calibrated specific protein structural properties per 10°C rise in living temperature of the organisms in a given family. The statistical trends obtained by Vogt *et al*. (1997) and by us are similar, indicating the equivalence of the two approaches.

The properties of these 18 pairs of thermophilic and mesophilic proteins are summarized in Table I. The best matching protein chains in each family are indicated in the footnotes of Table I. One PDB entry for the mesophilic protein elongation factor EF-TU-EF-TS complex (PDB entry 1EFU) from *Escherichia coli* is an A_2B_2 type tetramer with chains of type A and B being highly dissimilar. This particular protein complex has two different homologs in the thermophilic proteins, namely, EF-TU (PDB entry 1EFT) and EF-TS (PDB entry 1TFE). Furthermore, 1TFE, a dimer, matches with a single chain, 1EFU-B. The asymmetric unit of lactate dehydrogenase crystals from *Bacillus stearothermophilus* (PDB entry 1LDN) contains two copies of the molecule. The first copy has been used in this analysis. In all the families, the spatially overlapping regions in the superposition of the thermophilic and mesophilic proteins are very extensive. For example, in the citrate synthase family, where the similarity between the thermophilic and mesophilic proteins is relatively poor as compared with most other families, 332 residues in each chain overlap spatially. A chain of thermophilic citrate synthase (1AJ8-B) has 370 residues while a chain of mesophilic citrate synthase (1CSH) contains 435 residues. A few of the PDB entries used in this analysis have missing atoms, residues or small fragments due to poor diffraction data. Additionally, the crystal structures in several cases may be determined at low temperatures to obtain better diffraction data. However, these factors do not substantially affect the overall three-dimensional structures of the proteins. No systematic errors are expected on this count.

Sequence composition analysis

Distributions (numbers, *N*) and frequencies (percent, %) of all 20 amino acids were computed for the thermophilic and mesophilic proteins. In addition, we have computed their distributions in the α-helices. The amino acid distributions were compared using the χ^2-test. Hamming distance was computed between percent (%) amino acid compositions. The change in proportion test was used to identify the amino acids whose proportions change significantly. These calculations follow Kumar and Bansal (1998a).

Structural properties

Oligomeric state

For a given protein, the PDB files contain coordinates for the structure observed in a crystallographic asymmetric unit. This may not reflect the true biochemically relevant oligomeric state for the protein. In our data set these oligomeric states of the thermophilic and mesophilic proteins are tabulated by

studying the biochemical data contained in the relevant literature on these proteins, indicators within the PDB files and the pointers in the PDB3DB browser.

Hydrophobicity

The hydrophobicity of a protein was calculated as the fraction of the buried non-polar area out of the total non-polar area, computed by using the methods described earlier (Tsai and Nussinov, 1997a,b; Tsai *et al*., 1997).

Compactness

The compactness (Zehfus and Rose, 1986) of a protein was defined as the ratio of solvent accessible area (Lee and Richards, 1971; Tsai *et al*., 1997) of the protein and the surface area of a sphere with equal volume to the protein (Tsai and Nussinov, 1997a,b).

Hydrogen bonds and salt bridges

Whenever two heavy (non-hydrogen) atoms with opposite partial charges [donor (D)–accepter (A) pairs] were found to be within a distance of 3.5 Å, a hydrogen bond has been inferred. The geometrical goodness of the hydrogen bond was assessed by computing the values of the following angles.

- Angle θ_D between vectors **BD–D** and **D–A**, BD is the atom covalently bonded to the donor (D) atom.
- Angle θ_A between vectors **D–A** and **A–BA**, BA is the atom covalently bonded to the acceptor (A) atom.

A hydrogen bond was taken to have good geometry if both these angles lie in the range 90–150°. Only those hydrogen bonds which have a good geometry were included in our studies.

The presence of salt bridges was inferred when Asp or Glu side-chain carbonyl oxygen atoms were found to be within 4.0 Å distance from the nitrogen atoms in Arg, Lys and His side chains.

Helical content

The helical content of a protein refers to the percentage (%) of residues that have α-helical conformation in the protein. The corresponding Dictionary of Protein Secondary Structure (DSSP) (Kabsch and Sander, 1983) file was used to identify the residues in α-helical conformation in each protein. Overall geometries of α-helices in the thermophilic and mesophilic protein chains were characterized using HELANAL (Kumar and Bansal, 1996; Kumar and Bansal, 1998b). This program is available at http://www-lecb.ncifcrf.gov/~kumarsan/

Buried and exposed surface areas

Buried and accessible surface areas (Lee and Richards, 1971; Tsai and Nussinov, 1997a,b) have been computed for thermophilic and mesophilic protein chains as well as for 165 dissimilar monomers. Four different fractions have been computed from these areas, in each case:

- Fraction of polar exposed surface area is the ratio of the exposed polar surface area to the total exposed surface area.
- Fraction of non-polar exposed surface area is the ratio of the exposed non-polar surface area to the total exposed surface area.
- Fraction of polar buried surface area is the ratio of the buried polar surface area to the total buried surface area.
- Fraction of non-polar buried surface area is the ratio of the buried non-polar surface area to the total buried surface area.
- Total exposed surface area is the sum of polar and non-polar exposed surface areas. Similarly, the total buried surface area is the sum of polar and non-polar buried surface areas.

Measurement of percent change in various properties

For the purpose of a comparison between a thermophilic–mesophilic pair, the numbers of hydrogen bonds and salt bridges in the two proteins were normalized by their respective number of residues. Percent changes were computed as the difference between the normalized values of hydrogen bonds and salt bridges in the two proteins in each family, divided by the corresponding normalized values for the mesophilic proteins.

Changes in protein size can occur due to insertion/deletion and/or oligomerization. Percent change in protein size in each family was computed by dividing the difference in the number of residues between the thermophilic and mesophilic proteins by the number of residues in the mesophilic protein.

Percent change in hydrophobicity in each family was computed by dividing the difference in hydrophobicity for the thermophilic and mesophilic proteins by the hydrophobicity for the mesophilic protein. Percent change in compactness was also computed in the same way.

Database of 165 dissimilar monomers

A database of 165 proteins, which (i) have been solved to high resolution R $\leqslant$ 2.5 Å by X-ray crystallography and contain at least 50 amino acids, (ii) have dissimilar 3D structures, as determined by the sequence order independent structure comparison technique (Tsai *et al*., 1996), and (iii) exist as monomers in solution as indicated in their PDB files, relevant biochemical literature and pointers in PDB3DB browser to other databases such as SWISS-PROT, was generated from the PDB. This database was used as a control for studying structural features, such as compactness, hydrophobicity, polar and non-polar contribution to buried and exposed surfaces in thermophilic and mesophilic protein chains.

Cases of high resolution structural pairs where the melting temperatures are currently available

(i) 3-Phosphoglycerate kinase (PGK) (Davies *et al*., 1993): T_m = 67°C for the thermophilic enzyme from *Bacillus stearothermophilus* and 53°C for its mesophilic enzyme counterpart, from *Saccharomyces cerevisiae*. The thermophilic PGK is a monomer while the mesophilic PGK is a dimer. The energy difference between the two enzymes, $\Delta\Delta G$ = ~5 kcal/mol.

(ii) Adenylate kinase (Glaser *et al*., 1992): T_m = 74.5°C for the thermophilic enzyme from *Bacillus stearothermophilus* and 48°C for the mesophilic enzyme from *Saccharomyces cerevisiae*. Both the thermophilic and the mesophilic enzymes are monomers.

(iii) CheY, the bacterial chemotaxis protein (Usher *et al*., 1998): T_m for the thermophilic protein is 95°C from *Thermotoga maritima*. Both the thermophilic and the mesophilic proteins are monomers.

(iv) Glutamate dehydrogenase (Yip *et al*., 1995): T_m = 113°C for the thermophilic protein from *Pyrococcus furiosus*. Both the thermophilic and the mesophilic enzymes are hexamers. T_m = 55°C for *Clostridium symbiosum* glutamate dehydrogenase (Yip *et al*., 1995).

(v) Rubredoxin, a small redox protein (Day *et al*., 1992): there are several estimates of T_m for rubredoxin from *Pyrococcus furiosus*. The one used here is from Hiller *et al*. (1997), determined by the Hydrogen exchange technique. T_m for thermophilic rubredoxin = 176 – 195°C. Both the thermophilic and the mesophilic rubredoxins are monomers.

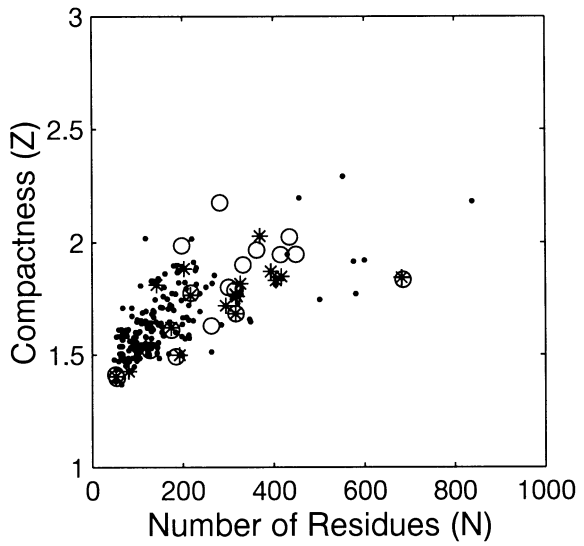

Fig. 1. Distribution of compactness as a function of chain size (number of residues), for thermophilic (*) and mesophilic (○) protein chains. *x*-axis denotes the number of residues (*N*) in the protein chains and *y*-axis denotes compactness (*Z*). For comparison, 165 monomers with dissimilar structures (●) obtained from the PDB are also depicted.

For PGK the melting temperatures of the thermophilic and mesophilic proteins are close (ΔT_m = 67 – 53 = 14°C). The energy difference between thermophilic and mesophilic enzymes is only 5 kcal/mol ($\Delta\Delta G$ = ~5 kcal/mol). Moreover, the oligomeric states of the two PGKs are also different. The thermophilic rubredoxin has a very high T_m. However, it is a very small protein, consisting of only about 50 amino acids. More than one estimate of T_m for rubredoxin further complicates the matter.

Results

We have selected a non-redundant dataset of 18 families consisting of thermophilic and mesophilic proteins whose high resolution ($R \leqslant 2.5$ Å) structures are available in the PDB (Table I). The corresponding thermophilic and mesophilic proteins within these families are highly similar, with sequence identities varying in a range of 24–73% and backbone r.m.s.d. values between 0.69 and 1.68 Å. At the same time, the thermophilic proteins across the 18 families are highly dissimilar among themselves (sequence identities being <10% and backbone r.m.s.d. > 2 Å). The mesophilic proteins are also highly dissimilar among themselves.

Packing

Reasons for higher stability of thermophilic proteins include better packing (Russell *et al*., 1997, 1998) and hence, smaller and less numerous cavities. To study packing in a protein one can compute its compactness (Zehfus and Rose, 1986). Compactness has been defined to be the ratio of accessible surface area (*ASA*) (Lee and Richards, 1971) of a given protein to the surface area of a sphere with the same volume as the protein. Assuming that most proteins are more or less globular in shape, a better packed protein will have a smaller ratio value. We have already used this formulation to study hydrophobic folding units (Tsai and Nussinov, 1997a,b). Figure 1 plots the compactness versus the number of residues in thermophilic and mesophilic protein chains (one chain per protein), along with the values calculated for the 165 structurally dissimilar monomeric protein chains selected from the PDB. The compactness values for the thermophilic protein chains are very similar to those calculated for the mesophilic protein chains. They are also within the range of the compactness values obtained for the 165 dissimilar monomers. However, the overall packing of an oligomeric protein may involve two components: (i) packing of atoms within individual subunits, and (ii) the association, or packing, of the subunits with respect to each other. Consequently, we have computed the compactness for the thermophilic and mesophilic proteins in their biochemically relevant oligomeric states. The results are presented in Table II. Again, the compactness values for thermophilic and mesophilic proteins are highly similar. Hence, there is no consistent pattern in the contribution of packing to the differences in stabilities between thermophilic and mesophilic protein pairs. Recently, Karshikoff and Ladenstein (1998) have also reached similar conclusions upon computing cavity volumes for a large number of thermophilic and mesophilic proteins.

Hydrophobicity

With the rapid increase in the structural information available for proteins, it is becoming increasingly clear that the hydrophobic effect is the dominant driving force in protein folding (Dill, 1990). Hence, it has been suggested that thermophilic proteins are substantially more hydrophobic (Haney *et al*., 1997) and have more surface area buried upon oligomerization (Salminen *et al*., 1996) as compared with their mesophilic counterparts. As with packing, the hydrophobic effect can manifest itself at two levels: (i) hydrophobicities of the individual protein chains, and (ii) hydrophobicity due to the association of the chains. We have computed the hydrophobicity as the fraction of buried non-polar surface area out of the total non-polar surface area (Tsai and Nussinov, 1997a,b), for the thermophilic and mesophilic protein chains as well as their biochemically relevant oligomeric forms. Figure 2 presents a plot of the hydrophobicity versus the number of residues in thermophilic and mesophilic protein chains, along with those for the 165 dissimilar monomeric chains. The figure illustrates that thermophilic and mesophilic protein chains have very similar hydrophobicities. The values lie within the same range as those for the hydrophobicities of 165 dissimilar monomers. The hydrophobicities computed for the thermophilic and mesophilic proteins in their biochemically relevant oligomeric states are presented in Table II. Again, the hydrophobicities of the thermophilic and mesophilic protein oligomers are very similar.

Polar and non-polar surface areas

It has been suggested that increased polar surface area contributes to the greater stability of the thermophilic proteins (Haney *et al*., 1997; Vogt and Argos, 1997; Vogt *et al*., 1997). Here, we have divided protein surfaces into buried and exposed parts and evaluated the contribution of polar and non-polar atoms. These calculations have been performed for all thermophilic and mesophilic protein chains (one polypeptide chain per protein) and compared with those for 165 dissimilar monomers. The calculations have been done in two different ways. In the first set all atoms including the backbone were considered. In the second set, the backbone atoms were excluded. Table III presents the results. The distributions of buried and exposed, polar and non-polar surface areas are quite uniform for the 165 dissimilar monomers as well as for the thermophilic and mesophilic protein chains.

Table II. Values of hydrophobicity and compactness in thermophilic and mesophilic proteins

Thermophilic proteins			Mesophilic proteins		
PDB entry	Hydrophobicity	Compactness	PDB entry	Hydrophobicity	Compactness
1AJ8 AB	87	1.952	1CSH AB	88	1.925
1BDM AB	85	2.033	4MDH AB	84	2.145
1CAA	75	1.403	8RXN	75	1.411
1CIU	88	1.843	1CDG	87	1.834
1EFT	84	1.831	1EFU C	82	1.966
1GTM A–F	88	2.721	1HRD A–F	88	2.765
1LDN A–D	89	2.119	1LDG A–D	88	2.090
1LNF EA	87	1.973	1NPC	84	1.683
1PHP	84	1.870	1QPG AB	85	2.271
1TFE AB	80	2.043	1EFU B	76	2.176
1TMY	79	1.529	3CHY	80	1.520
1XGS AB	86	1.969	1MAT	82	1.629
1YNA	83	1.499	1XNB	84	1.492
1ZIN	80	1.765	1AKY	80	1.775
2FXB	77	1.426	1FCA	74	1.396
2PRD A–F	87	2.156	1INO A–F	87	2.149
3MDS A–D	86	2.186	1QNM A–D	85	2.204
3PFK A–D	88	2.117	2PFK A–D	85	2.581

The values of hydrophobicity and compactness (Tsai and Nussinov, 1997a,b) presented here are for biochemically relevant oligomeric states of the thermophilic and mesophilic proteins. First four letters in columns for PDB entries denote four letter PDB code. Other letters represent protein subunits.

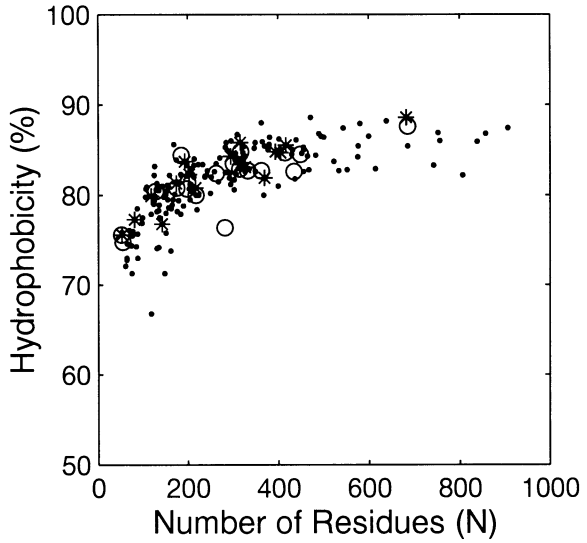

Fig. 2. Distribution of hydrophobicity as a function of chain size (number of residues), for thermophilic (*) and mesophilic (○) protein chains. *x*-axis denotes the number of residues (*N*) in the protein chains and *y*-axis denotes percent hydrophobicity. For comparison, 165 monomers with dissimilar structures (●) obtained from the PDB are also depicted.

Table III. Polar and non-polar contributions to buried and accessible surface areas

Fractional surface area	165 monomers	Thermophilic protein chains	Mesophilic protein chains
All atoms			
Frac-pol-exp-area	0.493 ± 0.036	0.489 ± 0.027	0.482 ± 0.031
Frac-nonpol-exp-area	0.507 ± 0.036	0.511 ± 0.027	0.518 ± 0.031
Frac-pol-buried-area	0.568 ± 0.014	0.557 ± 0.011	0.565 ± 0.013
Frac-nonpol-buried-area	0.432 ± 0.014	0.443 ± 0.011	0.435 ± 0.013
Side chain atoms			
Frac-pol-exp-area	0.476 ± 0.046	0.473 ± 0.032	0.464 ± 0.038
Frac-nonpol-exp-area	0.524 ± 0.046	0.527 ± 0.032	0.536 ± 0.038
Frac-pol-buried-area	0.254 ± 0.025	0.231 ± 0.029	0.236 ± 0.037
Frac-nonpol-buried-area	0.746 ± 0.025	0.769 ± 0.029	0.764 ± 0.037

Frac-pol-exp-area denotes the polar contribution to the exposed surface area. Frac-nonpol-exp-area denotes the nonpolar contribution to the exposed surface area. Frac-pol-buried-area denotes the polar contribution to the buried surface area. Frac-nonpol-buried-area denotes the nonpolar contribution to the buried surface area. These areas are defined in the Materials and methods. Thermophilic and mesophilic chains have similar polar and non-polar contributions to their buried and exposed surfaces.

The above observations on packing, hydrophobicity and surface areas indicate that basic protein core is similar between thermophiles and mesophiles.

Salt bridges and hydrogen bonds

Along with oligomerization, chain length, hydrophobicity and compactness, hydrogen bonds and salt bridges have also been compared between the thermophilic and the mesophilic proteins. The hydrogen bonds were divided into three classes: main chain–main chain (MM H-bonds), main chain–side chain (MS H-bonds) and side chain–side chain hydrogen bonds (SS H-bonds). Figure 3 shows plots of SS H-bonds and salt bridge content changes in the families of thermophilic and mesophilic proteins in their biochemically relevant oligomeric states, and at their interfaces. As the figure shows, side chain–side chain H-bonds and salt bridge content increase in the monomers of most thermophilic proteins and at their interfaces.

The most significant change in the number of salt bridges was observed in the glutamate dehydrogenase family. This family contains glutamate dehydrogenase enzymes from hyperthermophile *Pyrococcus furiosus* and the mesophile *Clostridium symbiosum*. Both thermophilic and mesophilic glutamate dehydrogenases are homohexamers and share good sequence and structural similarities (Table I). The difference between their melting temperatures is approximately 60° (see Materials and methods). *Pyrococcus furiosus* glutamate dehydrogenase contains 168 salt bridges while *Clostridium symbiosum* glutamate dehydrogenase contains 107 salt bridges. Thus, the salt

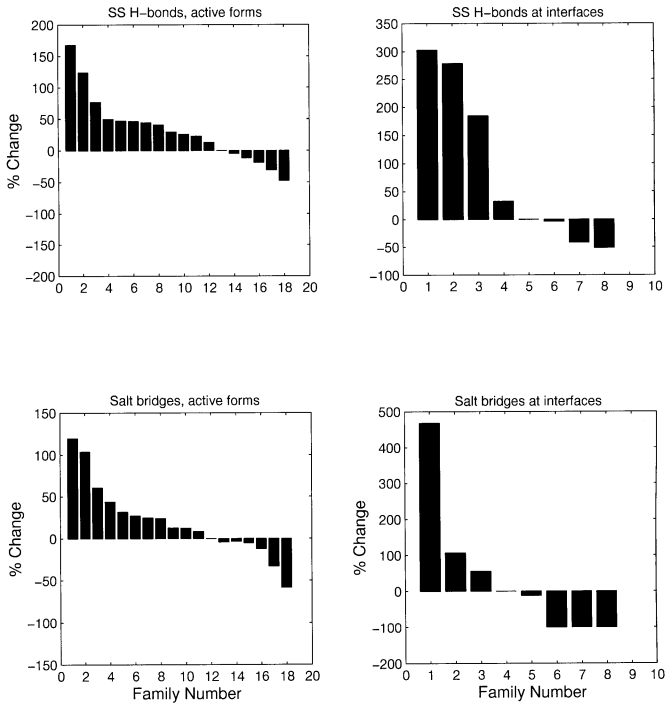

Fig. 3. Plots depicting changes in side chain–side chain hydrogen bonds (SS H-bonds) and salt bridges in biochemically relevant forms of proteins and at interfaces in various families of thermophilic and mesophilic proteins. A positive change indicates that the thermophilic protein has a higher content as compared with its mesophilic homolog, while a negative change indicates that the thermophilic protein has a lower content than its mesophilic homolog. For the majority of the families, SS H-bond and salt bridge content increases for thermophilic proteins. For each subplot, the *x*-axis denotes the family number while the *y*-axis represents the percent change in the property indicated at top of the subplot. The data on interfaces is available only in the case of eight families.

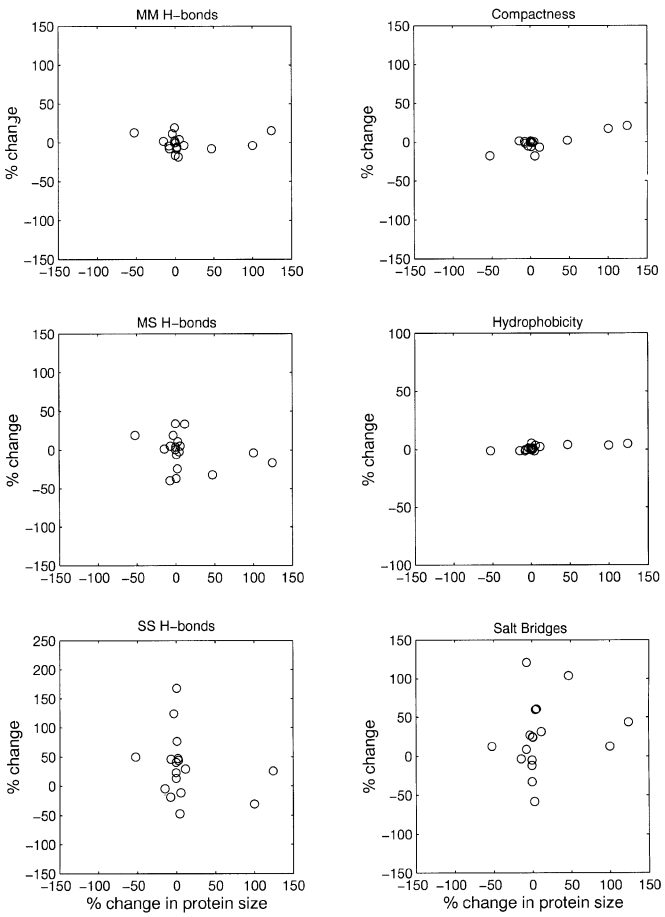

Fig. 4. Change in hydrogen bonds, salt bridges, compactness and hydrophobicity plotted with respect to change in protein size (number of residues in the protein). For each subplot, the *x*-axis denotes the percent change in the protein size and the *y*-axis represents the percent change in the property indicated at top of the subplot. Most structural properties of proteins are not correlated with insertion/deletion or oligomerization.

bridge frequency increases by ~70% for the thermophilic protein. The changes in other structural parameters between this thermophile–mesophile pair are insignificant (Table II; Yip *et al*., 1995). Thus salt bridges and their networks have been implicated in thermostability of this protein (Yip *et al*., 1995). Recently, we have computed the electrostatic strengths of salt bridges in monomers of this pair (Kumar *et al*., 2000). We have observed that salt bridges in *Pyrococcus furiosus* glutamate dehydrogenase, which form extensive networks, are highly stabilizing. On the other hand, salt bridges in *Clostridium symbiosum* glutamate dehydrogenase, which form considerably less networks, are only marginally stabilizing (Kumar *et al*., 2000).

Insertions, deletions and oligomerization

It has been suggested that deletion or shortening of loops may increase protein thermal stability (Russell *et al*., 1997, 1998). Oligomerization can be another contributing factor. These factors reflect a change in protein size, and its effect on thermal stability. Figure 4 shows changes in hydrogen bonds, salt bridges, compactness and hydrophobicity plotted against the change in the number of residues between thermophilic and mesophilic proteins in each family. Mostly there is no correlation with a change in protein size, either due to insertions/deletions or due to oligomerization. This is further corroborated by the observation that in 14 out of 18 families in our database, thermophilic and mesophilic proteins have the same oligomeric states. In two families the oligomeric states of thermophilic proteins are found to be higher than those of their mesophilic homologs. However, the oligomeric states of mesophilic proteins are higher than their thermophilic homologs in the other two families.

Living temperatures of the thermophilic organisms and structural factors involved in protein thermostability

In the literature, the stability of thermophilic proteins has been described in a number of ways, such as in terms of the temperature at which a protein is active (activity temperature), stable (stability temperature) or by half life for a certain duration of time. Much less frequently a protein is described in terms of melting, or mid-point transition temperature (T_m). Perhaps due to this heterogeneity in the available data, a recent database analysis study (Vogt and Argos, 1997; Vogt *et al*., 1997) used the living temperatures of the organisms from which the proteins were isolated as a parameter for studying thermostability. Figure 5 plots changes in the oligomeric state, chain length, hydrophobicity, compactness, main chain–main chain, main chain–side chain and side chain–side chain hydrogen bonds and salt bridges as a function of living and of melting temperatures. Figure 5a shows that structural factors involved in protein thermostability do not correlate with living temperatures of the thermophilic organisms. The trends

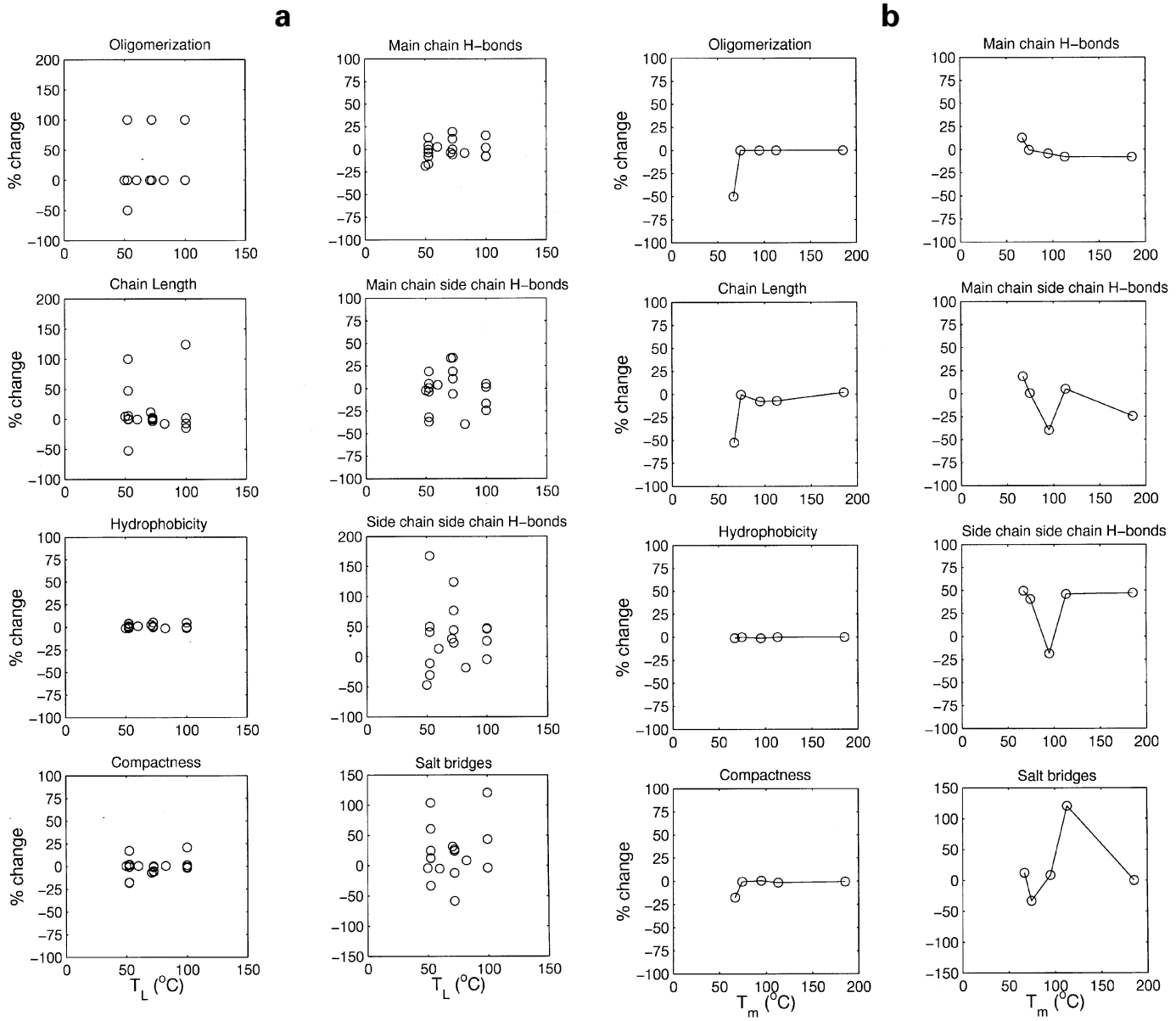

Fig. 5. Change in various structural properties—oligomerization, chain length, hydrophobicity, compactness, hydrogen bonds—involving main chain–main chain atoms, main chain–side chain atoms, side chain–side chain atoms and salt bridges plotted against (**a**) living temperature (T_L) of thermophilic organisms and (**b**) melting temperature (T_m) of the thermophilic proteins. Trends for various properties are clearer in the plots with T_m. Salt bridges show a correlation with melting temperature. However, the correlation is not statistically reliable. For each subplot, the *x*-axis represents the temperature, while the *y*-axis represents the percent change in property indicated at top of the subplot. In each panel of (b), the first point (smallest T_m) corresponds with phosphoglycerate kinase. The second point corresponds with adenylate kinase. The third point represents CheY. The fourth point corresponds with glutamate dehydrogenase and the fifth point (greatest T_m) represents rubredoxin.

observed in Figure 5b are clearer. However, there are only five data points, two out of these (first and last) are unreliable due to reasons summarized in the Materials and methods section. If we ignore these points, we observe that among the various factors, only the salt bridges tend to correlate with the melting temperature. Unfortunately, this observation is unreliable, as it is based only on three proteins. However, it is consistent with studies by Yip *et al.* (1998), who have observed a correlation between ion pairs and thermostability for glutamate dehydrogenases from different organisms. Clearly, this phenomenon needs to be investigated further before any conclusions are drawn.

Distribution of amino acids

The overall distributions of amino acids in the 18 non-redundant families of thermophilic and mesophilic protein chains are presented in Table IV. Figure 6 presents a comparison between the residue composition of the thermophilic and mesophilic proteins. Despite the high sequence homology, a χ^2 test (Kumar and Bansal, 1998a) indicates that the differences between the two distributions are highly significant ($\chi^2 = 86.2$). For a 19 parameter system such as amino acid distribution, a χ^2 value at 95% level of confidence (probability of accepting the null hypothesis that two distributions are similar, $P \leqslant 0.05$) should be greater than 30.14 to reject the null hypothesis.

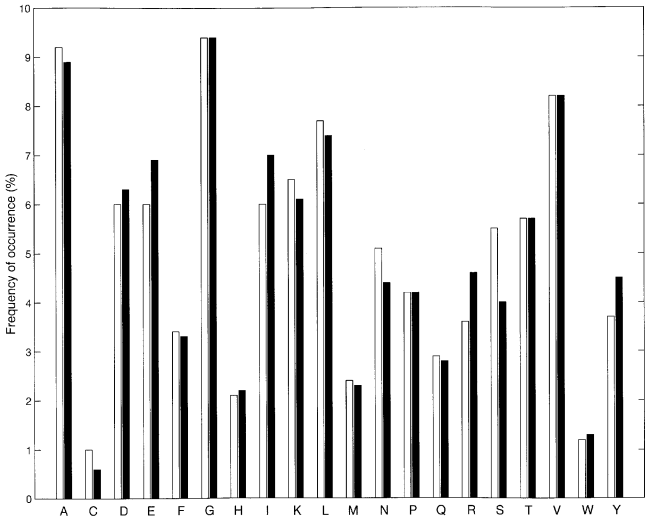

Fig. 6. Bar diagram showing a comparison between amino acid compositions of thermophilic and mesophilic protein chains. For each residue indicated by single letter code on the *x*-axis, the white bar represents frequency of occurrence (*y*-axis) of the residue in mesophilic protein chains and the black bar represents the same in thermophilic protein chains. Change in proportion tests show that differences in frequencies of Cys, Ser, Arg and Tyr are significant at a 95% level of confidence.

This evidence is further corroborated by the observation that the value of Hamming distance in 20 dimensional amino acid composition (%) space (Kumar and Bansal, 1998a) between thermophilic and mesophilic chains is large (8.1 distance units).

Proline substitutions

It has been suggested that Pro has an increased occurrence in thermophilic proteins, especially in loops (Haney *et al*., 1997; Watanabe *et al*., 1997; Bogin *et al*., 1998). A total of 75 Pro substitutions are observed in loop regions of thermophilic and mesophilic chains. In 39 cases, the thermophilic chains contain a Pro residue instead of other residues found in their mesophilic homologs at equivalent loop positions. However, in 36 cases, another residue is present in the thermophilic chains instead of Pro in the mesophilic homologs. Thus, there is no consistent pattern for Pro substitutions in loops. In our database, the frequency of occurrence of Pro is unchanged (4.2%) (Figure 6) in thermophilic and mesophilic proteins.

Preferred and avoided residues in thermophilic proteins

A change in proportion test (Kumar and Bansal, 1998a) is used to identify amino acids whose proportions change significantly, that is, by >2 standard deviations, between thermophilic and mesophilic chains. Changes in the proportions of Cys (0.6% in thermophilic and 1.0% in mesophilic chains), Arg (4.6% in thermophilic and 3.6% in mesophilic chains), Ser (4.0% in thermophilic and 5.5% in mesophilic chains) and Tyr (4.5% in thermophilic and 3.7% in mesophilic chains) are found to be significant (Figure 6).

Of the 20 amino acids, Asn, Gln, Met and Cys can be classified as thermolabile due to their tendency to undergo deamidation or oxidation at high temperatures (Russell *et al*., 1997). Table IV and Figure 6 indicate that the frequencies of occurrence for Gln (2.8% in thermophiles and 2.9% in mesophiles) and Met (2.3% in thermophiles and 2.4% in mesophiles) are similar. Cys (0.6% in thermophilic chains and 1.0% in mesophilic) and Asn (4.4% in thermophilic and 5.1% in mesophilic) change by appreciable amounts. However, only the change in the frequency of Cys is significant.

Table IV. Distribution of amino acid residues in the 18 non-redundant families of thermophilic and mesophilic proteins

Amino acid residue	Thermophilic proteins no. / (%)	Mesophilic proteins no. / (%)
Ala (A)	449 (8.9)	476 (9.2)
Cys (C)*	30 (0.6)	52 (1.0)
Asp (D)	316 (6.3)	313 (6.0)
Glu (E)	348 (6.9)	311 (6.0)
Phe (F)	167 (3.3)	178 (3.4)
Gly (G)	471 (9.4)	484 (9.4)
His (H)	108 (2.2)	111 (2.1)
Ile (I)	351 (7.0)	313 (6.0)
Lys (K)	304 (6.1)	338 (6.5)
Leu (L)	372 (7.4)	399 (7.7)
Met (M)	118 (2.3)	125 (2.4)
Asn (N)	219 (4.4)	262 (5.1)
Pro (P)	211 (4.2)	217 (4.2)
Gln (Q)	139 (2.8)	152 (2.9)
Arg (R)*	231 (4.6)	185 (3.6)
Ser (S)*	202 (4.0)	286 (5.5)
Thr (T)	285 (5.7)	297 (5.7)
Val (V)	412 (8.2)	422 (8.2)
Trp (W)	64 (1.3)	62 (1.2)
Tyr (Y)*	226 (4.5)	191 (3.7)
Total	5023	5174

No., the number of amino acid in the thermophilic and mesophilic protein chains.
%, percentage of amino acids in the thermophilic and mesophilic protein chains.
χ^2 value for the amino acid distributions in thermophilic and mesophilic proteins is 86.2, indicating that differences between them are highly significant. Hamming distance between the thermophilic and mesophilic proteins in 20-dimensional amino acid composition space is 8.1 distance units.
'*' identifies the amino acid residues whose proportions change significantly (>2 standard deviations) between the thermophilic and mesophilic proteins, as indicated by change of proportion test.

The above observations raise questions about the possible roles of Arg, Tyr and Ser whose proportions change significantly. It has been suggested that thermophilic proteins have increased hydrogen bonding and salt bridge formation (Yip *et al*., 1995; Querol *et al*., 1996; Vogt and Argos, 1997; Vogt *et al*., 1997; Russell *et al*., 1997, 1998). Due to their large side chains, Arg and Tyr may be useful both in short range local interactions and in long range interactions. The guanidium group in Arg can form salt bridges. On the other hand, due to its short side chain Ser forms mostly local interactions (Jeffrey and Saenger, 1991). Interestingly, it has recently been observed that hot spots for binding in protein interfaces are also rich in Arg, Tyr and Trp (Clackson and Wells, 1995; Bogan and Thorn, 1998). Hence, it appears that in both binding and folding at high temperatures, Arg and Tyr play a similar role, contributing toward protein stability. On the other hand, Trp occurs with a similar proportion in both thermophilic and mesophilic chains (Table IV and Figure 6). In contrast to Arg and Tyr, Trp is a hydrophobic residue with a bulky double ring side chain, usually occurring with low frequencies in proteins. Alternatively, it is possible that the absence of a noticeable trend for Trp, a rare residue, is due to its low counts in our sample.

Thermophilic and mesophilic α-helices

It has been suggested that thermophilic proteins have a higher helical content (Querol *et al*., 1996). In our database, we find that in nine out of the 18 families, thermophilic and mesophilic

Table V. Distribution of amino acid residues in the α-helices in thermophilic and mesophilic proteins

Amino acid residue	α-Helices in thermophilic proteins (no., %)	α-Helices in mesophilic proteins (no., %)
Ala (A)	226 (14.1)	176 (13.4)
Cys (C)*	2 (0.1)	11 (0.8)
Asp (D)	92 (5.7)	61 (4.6)
Glu (E)	142 (8.8)	102 (7.8)
Phe (F)	81 (5.0)	59 (4.5)
Gly (G)	66 (4.1)	60 (4.6)
His (H)*	32 (2.0)	44 (3.3)
Ile (I)	114 (7.1)	85 (6.5)
Lys (K)	124 (7.7)	99 (7.5)
Leu (L)	147 (9.1)	139 (10.6)
Met (M)	53 (3.3)	39 (3.0)
Asn (N)	51 (3.2)	48 (3.7)
Pro (P)	11 (0.7)	17 (1.3)
Gln (Q)	59 (3.7)	63 (4.8)
Arg (R)*	88 (5.5)	51 (3.9)
Ser (S)	62 (3.9)	50 (3.8)
Thr (T)	71 (4.4)	60 (4.6)
Val (V)	95 (5.9)	93 (7.1)
Trp (W)	19 (1.2)	13 (1.0)
Tyr (Y)	72 (4.5)	44 (3.3)
Total	1607	1314

No., number of an amino acid in α-helices in thermophilic and mesophilic proteins.
%, percentage of an amino acid in the α-helices in thermophilic and mesophilic proteins.
χ^2 value for the amino acid distributions in α-helices in thermophilic and mesophilic proteins is 52.2, indicating that differences between them are highly significant. Hamming distance between the α-helices in thermophilic and mesophilic proteins in 20-dimensional amino acid composition space is 15.1 distance units.
'*' identifies the amino acid residues whose proportions change significantly (>2 standard deviations) between the α-helices in thermophilic and mesophilic proteins, as indicated by change of proportion test.

chains have similar values for the fraction of residues in helical conformation (f_H), as identified using DSSP (Kabsch and Sander, 1983). However, on the whole, thermophilic proteins have a higher occurrence of residues in helical conformation. f_H for thermophilic chains is 32.0% as compared with 25.4% in the mesophilic chains. α-Helices in the thermophilic and mesophilic proteins adopt similar overall geometries as characterized using HELANAL (Kumar and Bansal, 1996; Kumar and Bansal, 1998b).

Tables V presents the amino acid distributions in α-helices of thermophilic and mesophilic chains. χ^2-test shows that amino acid distribution in α-helices of thermophilic proteins is significantly different from that of α-helices in mesophilic proteins. Hamming distance (Kumar and Bansal, 1998a) between the two distributions is 15.1 distance units in the 20 dimensional amino acid composition space. The proportions of Cys (0.1% in thermophilic and 0.8% in mesophilic helices), His (2.0% in thermophilic and 3.3% in mesophilic helices) and Arg (5.5% in thermophilic and 3.9% in mesophilic helices) change significantly. Thermophilic helices favor Arg and avoid His and Cys as compared with mesophilic helices. A recent database analysis study on α-helices shows Arg to be a helix-favoring residue with its propensity to occur in the middle region of α-helices being 1.33, while Cys (propensity = 0.87 in the middle of α-helices) and His (propensity = 0.76 in the middle of α-helices) are helix disfavoring residues (Kumar and Bansal, 1998a). Thermostability has also been attributed to enhanced secondary structure propensity (Querol *et al*., 1996). This might rationalize the increase in the proportion of Arg, a helix favoring residue in thermophilic protein helices, while helix disfavoring residues Cys and His decrease. A previous analysis of the composition of α-helices in the thermophilic proteins (Warren and Petsko, 1995) has also noted a significant decrease in Cys and His. The proportion of Arg increases and that of Cys decreases significantly in the entire thermophilic proteins as well. Furthermore, Proline occurs with a frequency of 0.7% in α-helices of thermophilic as compared to 1.3% in α-helices of mesophilic proteins. Proline is the most avoided residue in the middle of α-helices (Kumar and Bansal, 1998a), since it may cause kinks (Woolfson and Williams, 1990; Kumar and Bansal, 1996, 1998a,b).

From the sequence composition comparison between thermophiles and mesophiles, thermophiles favor those factors that can enhance their stability, and avoid those factors which can destabilize them. Lower occurrence of thermolabile residues in the thermophilic chains along with lower occurrence of Cys, His and Pro in thermophilic helices illustrate a clear trend in this direction.

Discussion and conclusions

In this extensive study we have examined structural and sequence factors involved in protein thermostability. Thermophilic proteins optimize their stabilities via different mechanisms. Sequence and structural factors, such as packing, oligomerization, insertions and deletions, proline substitutions, helical content, helical propensities, polar surface area, hydrogen bonds and salt bridges, have been proposed to contribute to greater stability of thermophilic proteins. We have analyzed all these factors in a database of 18 thermophile–mesophile families. There are two major concerns in the analyses such as the ones presented here. First, protein stabilization strategies that may be observed in the individual families may not show consistent trends across several families. Second, not all differences among the thermophiles and mesophiles may be attributable to protein thermostability. Some may be due to phylogenetic differences between the thermophiles and mesophiles. In the available data, we observe that no single factor proposed to contribute toward protein thermostability is 100% consistent in our set of proteins. It is particularly interesting to note that hydrophobicity, packing and fractional polar and non-polar surface areas show little quantitative differences between thermophiles and mesophiles. While insertions/deletions, oligomerization and proline substitutions can stabilize individual thermophilic proteins, they do not show consistent trends across the families. It is also possible that the observed differences are due to phylogenetic differences between thermophiles and mesophiles. It should also be mentioned that more than one factor may be responsible for greater stability of the thermophilic protein in a given family.

The most consistent trend is shown by salt bridges and side chain–side chain hydrogen bonds. These increase in the majority of the thermophilic proteins. In recent years, the role of salt bridges toward protein stability has been controversial (Hendsch and Tidor, 1994; Kumar and Nussinov, 1999). However, in the case of the thermophilic proteins, salt bridges have been shown to be stabilizing (Elcock, 1998; Xiao and Honig, 1999; Kumar *et al*., 2000). Recently, we have calculated the electrostatic strengths of salt bridges in the glutamate dehydrogenase family (Kumar *et al*., 2000). Network formation stabilizes individual salt bridges in *Pyrococcus furiosus* glutam-

ate dehydrogenase (Kumar *et al*., 2000). Salt bridges are major contributors toward thermostability of *Pyrococcus furiosus* glutamate dehydrogenase as compared with the mesophilic *Clostridium symbiosum* glutamate dehydrogenase (Yip *et al*., 1995). In a large database analysis study, we have observed that salt bridges with 'good geometries', such as those in the present study, have mostly, but not always, contributed stabilizing electrostatic contributions toward protein stability (Kumar and Nussinov, 1999). Thermophilic proteins are not only stable, but are also optimally active at high temperatures. An increase in the number of salt bridges and hydrogen bonds may rigidify a thermophilic protein and expose it to the danger of becoming inactive. Still, while a thermophilic protein may be rigid at room temperature, it is likely to be flexible at high temperatures (Jaenicke and Bohm, 1998). Recently, we have also observed that *Pyrococcus furiosus* glutamate dehydrogenase contains a greater number of salt bridges and their networks around the active site as compared with the mesophilic *Clostridium symbiosum* glutamate dehydrogenase. The salt bridges around the active site may help to keep the active site region together by opposing disorder due to greater atomic mobility at high temperatures (Kumar *et al*., 2000).

Examination of the sequences shows that despite high sequence homology, the differences in amino acid distributions in the thermophilic and mesophilic proteins are highly significant. While some of the differences in the amino acid distributions are likely to be the outcome of phylogenetic differences between thermophiles and mesophiles, others correlate with protein thermostability. For example, the proportions of the thermolabile amino acid Cys, and of Ser which usually forms local interactions, decrease significantly, while those of Arg and Tyr which are capable of both short range and long range interactions increase significantly in the thermophilic proteins. The stability of the constituent α-helices also appears to contribute to protein thermal stability. Thermophilic proteins have a higher proportion of residues in helical conformation. Helix-favoring residue Arg occurs more frequently in α-helices of thermophilic proteins, whereas helix-disfavoring residues Cys, His and Pro have lower frequencies of occurrence in thermophilic helices. Refraining from using some residues, and opting for others in sequences of thermophilic proteins suggests a dual strategy employed by these proteins to enhance their stability. On the one hand, thermophilic proteins prefer residues with larger side chains that can form salt bridges, long range or local electrostatic and hydrophobic interactions, and which stabilize secondary structure elements. However, concomitantly, thermophilic proteins avoid thermolabile residues and residues that can destabilize secondary structure elements.

Our analysis shows that the organisms' living temperatures are not good descriptors of protein thermostability. Melting temperatures may be more appropriate to measure protein thermostability. When explored with respect to the melting temperatures, salt bridges appear to show a correlation with the T_m's. We note, however, that while high quality crystal structures are available, unfortunately, the T_m's have been determined only for a few of these proteins. Hence, currently we are unable to examine a correlation of salt bridges and the respective melting temperatures of the thermophiles in a statistically meaningful way. However, we observe that structural factors involved in the stability of the thermophilic proteins do not correlate with the living temperatures of their source organisms.

From the point of view of designing a thermophilic protein, this study suggests inclusion of a larger proportion of salt bridges. Additionally, it indicates including residues in α-helical conformation, and a higher frequency of Arg both to form salt bridges and additionally to stabilize α-helices. It would be preferable to avoid Pro, Cys and His in α-helices, and avoid thermolabile residues, particularly Cys.

Acknowledgements

We thank Drs Buyong Ma and Neeti Sinha and, in particular, Dr Jacob V.Maizel for helpful discussions. The personnel at FCRDC are thanked for their assistance. The research of R.Nussinov in Israel has been supported in part by grant no. 95-00208 from BSF, Israel, by a grant from the Ministry of Science, by the Center of Excellence, administered by the Israel Academy of Sciences, by the Magnet grant, and by the Tel Aviv University Basic Research and Adams Brain Center grants. This project has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, under contract No. NO1-CO-56000. The content of this publication does not necessarily reflect the view or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products or organization imply endorsement by the U.S. Government.

References

Adams,M.W.W. and Kelly,R.M. (1995) *Chem. Engng News*, **73**, 32–42.
Auerbach,G., Jacob,U., Grottinger,M., Schurig,M. and Jaenicke,R. (1997) *Biol. Chem.*, **378**, 327–329.
Bernstein,F., Koetzle,T., Williams,G., Meyer,E.J., Brice,M., Rodgers,J., Kennard,O. Shimanuchi,T. and Tasumi,M. (1977) *J. Mol. Biol.*, **112**, 535–542.
Bogan,A.A. and Thorn,K.S. (1998) *J. Mol. Biol.*, **280**, 1–9.
Bogin,O., Peretz,M., Hacham,Y., Korkhin,Y., Frolow,F., Kalb(Gilboa),A.J. and Burstein,Y. (1998) *Protein Sci.*, **7**, 1156–1163.
Clackson,T. and Wells,J.A. (1995) *Science*, **267**, 383–386.
Daniel,R.M., Cowan,D.A., Morgan,H.W. and Curran,M.P. (1982) *Biochem. J.*, **207**, 641–644.
Davies,G.J., Gamblin,S.J., Littlechild,J.A. and Watson,H.C. (1993) *Proteins*, **15**, 283–289.
Day,M.W., Hsu,B.T., Joshua-Tor,L., Park,J.B., Zhou,Z.H., Adams,M.W.W. and Rees,D.C. (1992) *Protein Sci.*, **1**, 1494–1507.
Dill,K.A. (1990) *Biochemistry*, **31**, 7134–7155.
Elcock,A.H. (1998) *J. Mol. Biol.*, **284**, 489–502.
Fukuyama,K., Nagahara,Y., Tsukihara,T., Katsube,Y., Hase,T. and Matsubara,H. (1988) *J. Mol. Biol.*, **199**, 183–193.
Glaser,P., Presecan,E., Delepierre,M., Surewicz,W.K., Mantsch,H.H., Barzu,O. and Giles,A.M. (1992) *Biochemistry*, **31**, 3038–3043.
Gomes,J., Gomes,I., Kreiner,W., Esterbauer,H., Sinner,M. and Steiner,W. (1993) *J. Biotech.*, **30**, 283–297.
Haney,P., Konisky,J., Koretke,K.K., Luthey-Schulten,Z. and Wolynes,P.G. (1997) *Proteins*, **28**, 117–130.
Hendsch,Z.S. and Tidor,B. (1994) *Protein Sci.*, **3**, 211–226.
Hiller,R., Zhou,Z.H., Adams,M.W.W. and Englander,S.W. (1997) *Proc. Natl Acad. Sci. USA*, **94**, 11329–11332.
Holland,D.R., Hausrath,A.C., Juers,D. and Matthews,B.W. (1995) *Protein Sci.*, **4**, 1955–1965.
Jaenicke,R. and Bohm,G. (1998) *Curr. Opin. Struct. Biol.*, **8**, 738–748.
Jeffrey,G.A. and Saenger,W. (1991) *Hydrogen Bonding in Biological Structures*. Springer-Verlag, Berlin
Jiang,Y., Nock,S., Nesper,M., Sprinzl,M. and Sigler,P.B. (1996) *Biochemistry*, **35**, 10269–10278.
Kabsch,W. and Sander,C. (1983) *Biopolymers*, **22**, 2577–2637.
Karshikoff,A. and Ladenstein,R. (1998) *Protein Engng*, **1**, 867–872.
Kelly,C.A., Nishiyama,M., Ohnishi,Y., Beppu,T. and Birktoft,J.J. (1993) *Biochemistry*, **32**, 3913–3922.
Kjeldgaard,M., Nissen,P., Thirup,S. and Nyborg,J. (1993) *Structure*, **1**, 35–50.
Klump,H.H., Dikuggiero,J., Kessel,M., Park,J.B., Adams,M.W.W. and Robb,F.T. (1992) *J. Biol. Chem.*, **267**, 22681–22685.
Knegtel,R.M.A., Wind,R.D., Rozeboom,H.J., Kalk,K.H., Buitelaar,R.M., Dijkhuizen,L. and Dijkstra,B.W. (1996) *J. Mol. Biol.*, **256**, 611–622.
Kumar,S. and Bansal,M. (1996) *Biophys. J.*, **71**, 1574–1586.
Kumar,S. and Bansal,M. (1998a) *Proteins*, **31**, 460–476.
Kumar,S. and Bansal,M. (1998b) *Biophys. J.*, **75**, 1935–1944.
Kumar,S. and Nussinov,R. (1999) *J. Mol. Biol.*, **293**, 1241–1255.
Kumar,S., Ma,B., Tsai,C.J. and Nussinov,R. (2000) *Proteins*, **38**, 368–383.

Ladenstein,R. and Antranikian,G. (1998) *Adv. Biochem. Engng Biotechnol*., **61**, 37–85.
Lee,B.K. and Richards,F.M. (1971) *J. Mol. Biol.*, **55**, 379–400.
Matthews,B.W., Weaver,L.H. and Kester,W.H. (1974) *J. Biol. Chem.*, **249**, 8030–8044.
Obmolova,G., Kuranova,I. and Teplyakov,A. (1993) *J. Mol. Biol.*, **232**, 312–313.
Perutz,M. and Raidt,H. (1975) *Nature*, **255**, 256–259.
Querol,E., Perez-Pons,J.A. and Mozo-Villarias,A. (1996) *Protein Engng*, **9**, 256–271.
Russell,R.J.M., Ferguson,J.M.C., Haugh,D.W., Danson,M.J. and Taylor,G.L. (1997) *Biochemistry*, **36**, 9983–9994.
Russell,R.J.M., Gerike,U., Danson,M.J., Hough,D.W. and Taylor,G.L. (1998) *Structure*, **6**, 351–361.
Rypniewski,W.R. and Evans,P.R. (1989) *J. Mol. Biol.*, **207**, 805–821.
Salminen,T., Teplyakov,A., Kankare,J., Cooperman,B.S., Lahti,R. and Goldman,A. (1996) *Protein Sci.*, **5**, 1014–1025.
Singleton,P. and Sainsbury,D. (1978) *Dictionary of Microbiology and Molecular Biology*, 2nd Edn. John Wiley, New York.
Tsai,C.J., Lin,S.L., Wolfson,H. and Nussinov,R. (1996) *J. Mol. Biol.*, **260**, 604–620.
Tsai,C.J. and Nussinov,R. (1997a) *Protein Sci.*, **6**, 24–42.
Tsai,C.J. and Nussinov,R. (1997b) *Protein Sci.*, **6**, 1426–1437.
Tsai,C.J., Xu,D. and Nussinov,R. (1997) *Protein Sci.*, **6**, 1–13.
Tsunasawa,S., Izu,Y., Miyagi,M. and Kato,I. (1997) *J. Biochem.*, **122**, 843–850.
Usher,K.C., De la Cruz,A.F.A., Dahlquist,F.A., Swanson,R.V., Simon,M.I. and Remington,S.J. (1998) *Protein Sci.*, **7**, 403–412.
Vogt,G. and Argos,P. (1997) *Fold. Des*., **2**, S40–S46.
Vogt,G., Woell,S. and Argos,P. (1997) *J. Mol. Biol.*, **269**, 631–643.
Warren,G.L. and Petsko,G.A. (1995) *Protein Engng*, **8**, 905–913.
Watanabe,K., Hata,Y., Kizaki,H., Katsube,Y. and Suzuki,Y. (1997) *J. Mol. Biol.*, **269**, 142–153.
Wigley,D.B., Gamblin,S.J., Turkenburg,J.P., Dodson,E.J., Piontek,K., Muirhead,H. and Holbrook,J.J. (1992) *J. Mol. Biol.*, **223**, 317–335.
Woolfson,D.N. and Williams,D.H. (1990) *FEBS Lett.*, **277**, 185–188.
Xiao,L. and HonigB. (1999) *J. Mol. Biol.*, **289**, 1435–1444.
Yip,K.S.P. *et al*. (1995) *Structure*, **3**, 1147–1158.
Yip,K.S.P., Britton,K.L., Stillman,T.J., Lebbink,J., De Vos,W.M., Robb,F.T., Vetriani,C., Maeder,D. and Rice,D.W. (1998) *Eur. J. Biochem.*, **255**, 336–346.
Zehfus,M.H. and Rose,G.D. (1986) *Biochemistry*, **25**, 5759–5765.
Zuber,H. (1988) *Biophys. Chem*., **29**, 171–179.

Received June 30, 1999; revised October 26, 1999; accepted November 29, 1999