Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits: results of a comprehensive survey

András Szilágyi[1,2] and Péter Závodszky[1,2]*

Background: Proteins from thermophilic organisms usually show high intrinsic thermal stability but have structures that are very similar to their mesophilic homologues. From prevous studies it is difficult to draw general conclusions about the structural features underlying the increased thermal stability of thermophilic proteins.

Results: In order to reveal the general evolutionary strategy for changing the heat stability of proteins, a non-redundant data set was compiled comprising all high-quality structures of thermophilic proteins and their mesophilic homologues from the Protein Data Bank. The selection (quality) criteria were met by 64 mesophilic and 29 thermophilic protein subunits, representing 25 protein families. From the atomic coordinates, 13 structural parameters were calculated, compared and evaluated using statistical methods. This study is distinguished from earlier ones by the strict quality control of the structures used and the size of the data set.

Conclusions: Different protein families adapt to higher temperatures by different sets of structural devices. Regarding the structural parameters, the only generally observed rule is an increase in the number of ion pairs with increasing growth temperature. Other parameters show just a trend, whereas the number of hydrogen bonds and the polarity of buried surfaces exhibit no clear-cut tendency to change with growth temperature. Proteins from extreme thermophiles are stabilized in different ways to moderately thermophilic ones. The preferences of these two groups are different with regards to the number of ion pairs, the number of cavities, the polarity of exposed surface and the secondary structural composition.

Addresses: [1]Institute of Enzymology, Biological Research Centre, Hungarian Academy of Sciences, H-1518 Pf. 7 Budapest, Hungary and [2]Department of Biological Physics, Eötvös Loránd University, Pázmány Péter stny 1/A, H-1117 Budapest, Hungary.

*Corresponding author.
E-mail: zxp@enzim.hu

Key words: hyperthermophiles, ion pairs, protein structure, thermophiles, thermostability

Received: **15 January 2000**
Revisions requested: **10 February 2000**
Revisions received: **25 February 2000**
Accepted: **28 February 2000**

Published: **26 April 2000**

Structure 2000, **8**:493–504

0969-2126/00/$ – see front matter

Introduction

Proteins come in a wide variety of shapes and folds and possess a wide range of thermal stabilities. Proteins from thermophilic organisms usually exhibit substantially higher intrinsic thermal stabilities than their counterparts from mesophilic organisms [1,2] while retaining the basic fold characteristic of the particular protein family. Although the molecular underpinnings of protein thermostabilization have been the focus of many theoretical and experimental research efforts (for reviews, see [1–5]), this subject is only partially understood. Studies of thermostability can be divided into two categories: those examining a single thermophilic protein, comparing its atomic structure with one or more mesophilic homologues; and mostly computational studies that utilize data for a group of proteins, analyzing various features systematically in order to reach general conclusions. There are numerous examples for the first category [6–16].

The number of systematic studies is much smaller. Argos and colleagues compared the sequences of mesophilic and thermophilic proteins in three [17] and later six [18] protein families; in the latter study some three-dimensional structures were also considered. The main results of these studies were the 'traffic rules' for preferred amino acid exchanges between mesophilic and thermophilic proteins; these exchanges were thought to increase helical propensities and the hydrophilicity of the exposed surface as well as to decrease overall flexibility of the polypeptide chain. Further developments, however, indicated that the sample sets used in these studies were too small: many subsequently described thermophilic proteins did not follow the 'rules' [19].

Spassov and colleagues [20] introduced parameters to evaluate the degree of optimization of hydrophobic and charge–charge interactions in protein structures; their study involved 14 thermophilic protein structures. They concluded that these proteins are characterized by a higher degree of hydrophobic or electrostatic optimization than mesophilic ones. Warren and Petsko [21] compared the amino acid composition of α helices of 19 thermophilic proteins with the known average composition for mesophilic helices and found significant shifts in the probabilities for some amino acids.

Karshikoff and Ladenstein [22] studied the role of packing density in thermostability by examining 24 thermophilic proteins; 16 of them were compared with their mesophilic homologues. The authors concluded that mesophilic and thermophilic proteins essentially do not differ in the degree of packing. Argos and colleagues [23,24] created a data set containing 19 thermophilic and 37 related mesophilic protein structures and analyzed them for the number and type of hydrogen bonds and salt links, polar surface composition, internal cavities and packing densities, as well as secondary structure composition. They also correlated these properties with the growth temperatures of the source organisms of the proteins in their data set. The main conclusion from this work was that internal hydrogen bonds and salt bridges show a clear increase with increased thermostability; in addition, the polar surface fraction and secondary structure content also show some increase.

From this diverse collection of studies, it is difficult to draw a general conclusion about the structural features underlying the increased thermal stability of proteins from thermophilic microorganisms. The contradictions and this limited understanding is, in our judgment, a consequence of the limited amount of data and the non-uniform approach of the contributing researchers.

The rapidly growing number of structures in the Protein Data Bank (PDB) prompted us to devote a new study to the problem of protein thermostability, extending the analysis to all available structures and including a number of important properties that can be calculated from the atomic coordinates. In this work, we constructed a high-quality, non-redundant data set containing all available high-quality structures of thermophilic proteins that have mesophilic homologues with structures of equally high quality. The data set, based on the November 1998 version of the PDB, contains 29 thermophilic and 64 mesophilic protein subunit structures in 25 protein structural families and is the largest such data set created so far. For each structure in the data set, we determined 13 properties in five categories (cavities, hydrogen bonds, ion pairs, secondary structure and polarity of surfaces) from the atomic coordinates and performed a comparison of properties calculated for mesophilic and thermophilic structures using the methods of statistical analysis. In this study, the subunits are considered ignoring all interactions with other subunits in oligomeric proteins. As subunit–subunit interactions might have special importance in thermostability, a separate study will be devoted to them.

Results

Construction of a data set

For this work, a data set containing homologous mesophilic and thermophilic protein structures was constructed. There were four main requirements for the data set:

Firstly, the data set should be complete (i.e., it should contain all available structures from the PDB that satisfy the following requirements). We intended to use all available data for our analysis.

Secondly, the data set should be non-redundant and representative (i.e., it should contain only one PDB entry for any given protein as defined by its amino acid sequence). A given protein from a given organism is usually found in several PDB entries as a particular protein might have been crystallized in different forms and under varying conditions (e.g. point-mutated forms, forms with different bound ligands, at different temperatures and under different solvent conditions, etc.); in addition, many proteins consist of identical subunits. These structures are essentially identical, however, and using them all in our data set would have made it highly redundant. In these cases, a single representative structure should be selected and included in the data set.

Thirdly, the proteins in the data set should be divided into structural families. Each family should contain at least one thermophilic and at least one mesophilic structure for comparison. Comparisons are made between mesophilic and thermophilic structures within the same family. Low-temperature structures are not to be compared with room-temperature structures because low temperature might alter the structure significantly.

Fourthly, the structures in the data set should be of high quality. Missing chain pieces or sidechains and overall poor quality make the structure unsuitable for comparisons.

To meet these requirements, a procedure was designed to create the data set (see the Materials and methods section). In the final data set, there are 25 protein families with 64 mesophilic and 29 thermophilic protein subunits.

Based on the results of our calculations, we found that it is reasonable to distinguish between proteins from organisms with an optimum growth temperature (T_{opt}) between 45°C and 80°C and those with a $T_{opt} \approx 100$°C. The latter subset, containing only five subunit structures (PDB accession codes 1aisB, 1gtmA, 1caa, 1pczA and 1a1s), was designated as S_{100} and the former, containing the remaining 24 thermophilic subunits, was designated as $S_{45–80}$. There are no protein structures from organisms with a T_{opt} between 80°C and 100°C. For the sake of simplicity, we will refer to these subsets as 'moderately thermophilic proteins' ($S_{45–80}$) and 'extremely thermophilic proteins' (S_{100}), respectively, although in the literature organisms with T_{opt} around 80°C are often called 'extremely thermophilic'.

The statistical analysis of the differences between thermophilic structures and their mesophilic homologues was performed for the whole data set and also separately for

Table 1

Important parameters resulting from the statistical analysis of the data calculated from the protein structures in our data set.

Property	r_{full}	$r_{45–80}$	P_{full}	$P_{45–80}$	P_{100}
CavN	−0.352	0.095	−, 0.186	+, 0.378	−, 0.033
CavV	−0.154	0.071	+, 0.243	+, 0.129	−, 0.172
CavA	−0.221	0.069	+, 0.320	+, 0.127	−, 0.102
HboN	0.061	0.121	+, 0.324	+, 0.326	+, 0.450
UhbN	−0.160	−0.244	−, 0.305	−, 0.349	−, 0.216
Ip4N	0.323	0.135	+, 0.043	+, 0.178	+, 0.026
Ip6N	0.387	0.012	+, 0.0022	+, 0.024	+, 0.016
Ip8N	0.566	0.333	+, 0.00044	+, 0.0048	+, 0.014
HelC	0.137	0.429	+, 0.200	+, 0.152	−, 0.433
BetC	0.281	0.067	+, 0.281	−, 0.443	+, 0.052
IrrC	−0.272	−0.338	−, 0.168	−, 0.263	−, 0.187
ExPA	−0.446	−0.447	+, 0.127	+, 0.024	−, 0.278
BuPA	0.158	−0.213	+, 0.047	+, 0.156	+, 0.101

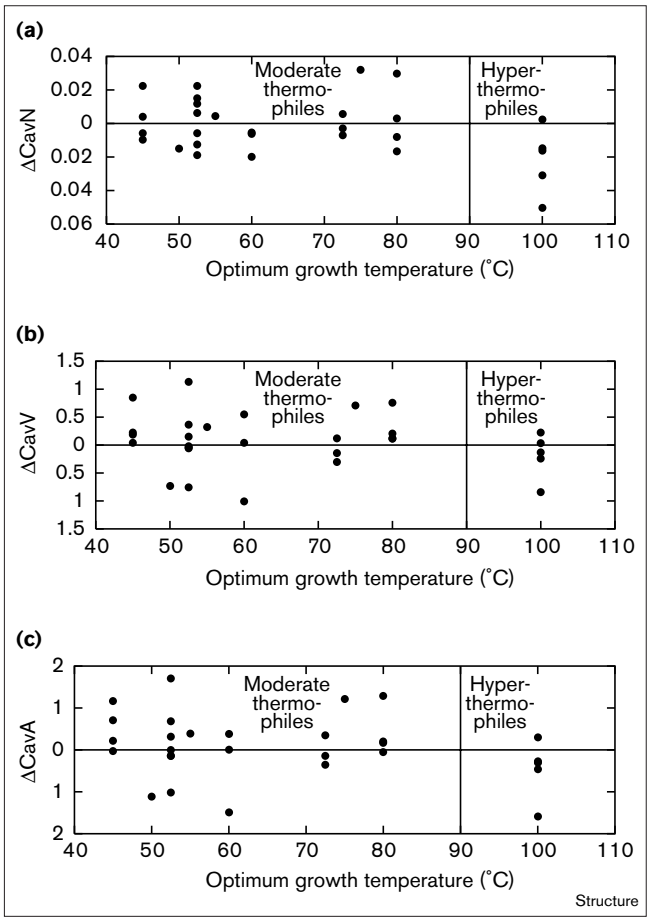
See text for explanations of property abbreviations. The second and third columns show the correlation coefficients (r) between T_{opt} and the difference between a property calculated for a thermophilic structure and the average of its mesophilic homologues. The sign of the averaged differences (+ or –) is shown in the fourth, fifth and sixth columns, along with the P values obtained from t tests described in the text. Indexes 'full', '45–80' and '100' refer to parameters calculated from the full data set, or the $S_{45–80}$ or S_{100} subset, respectively.

$S_{45–80}$ and S_{100}, in order to quantitatively characterize the different stabilization strategies of each group. Table 1 shows the correlation coefficients between the temperature and the differences between the thermophilic structures and their mesophilic homologues as well as the sign of the averaged differences and the P values obtained from the t tests described in the Materials and methods section. In these t tests, the null hypothesis was that there is no difference in a given property between thermophilic and mesophilic structures; the P value obtained from a t test is the probability that the observed data have arisen assuming that the null hypothesis is true. Thus, the smaller the P value the more statistically significant the difference between the two sets and, therefore, the less probable it is that there is actually no difference between thermophilic and mesophilic structures in the given property.

Cavities

Figure 1 shows the differences between the normalized number (CavN), total volume (CavV) and total surface area (CavA) of cavities of each thermophilic subunit and the average of its mesophilic homologues as a function of the optimum growth temperature of the source organism. The first three rows in Table 1 show the correlation coefficients, signs of averaged differences (+ or –) and P values for CavN, CavV and CavA for the full data set and the subsets. It is immediately clear that there is great variance in the data and extremely thermophilic proteins differ markedly from moderately thermophilic ones. Although none of the three parameters shows any significant difference between mesophilic and thermophilic proteins in

Figure 1



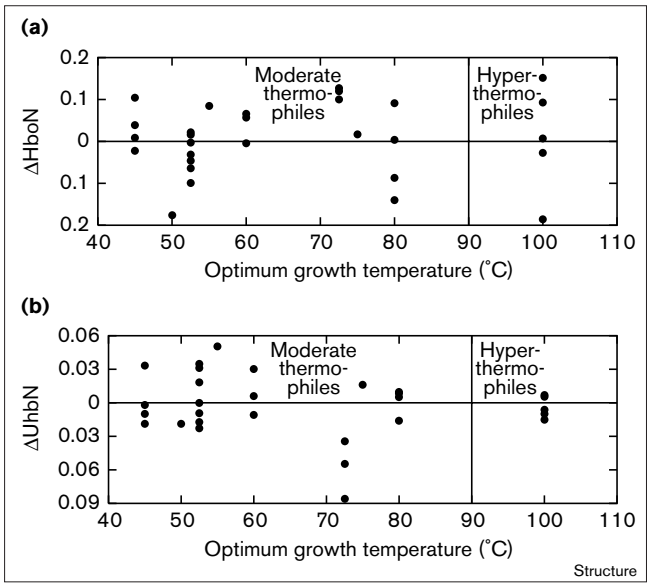
Differences between **(a)** the normalized number, **(b)** the total volume and **(c)** the total surface area of cavities of each thermophilic subunit and the average of its mesophilic homologues as a function of the optimum growth temperature of the thermophilic source organism.

$S_{45–80}$, the total surface area and especially the number of cavities show a marked decrease in the thermophilic proteins in S_{100}. The correlation coefficients between the temperature and the differences also reflect this situation: $r_{45–80}$ (calculated for $S_{45–80}$) is very close to zero for all the three parameters but r_{full} (calculated for the full data set) is relatively large and negative, indicating that there is a strong decrease in the parameters towards 100°C.

Hydrogen bonds

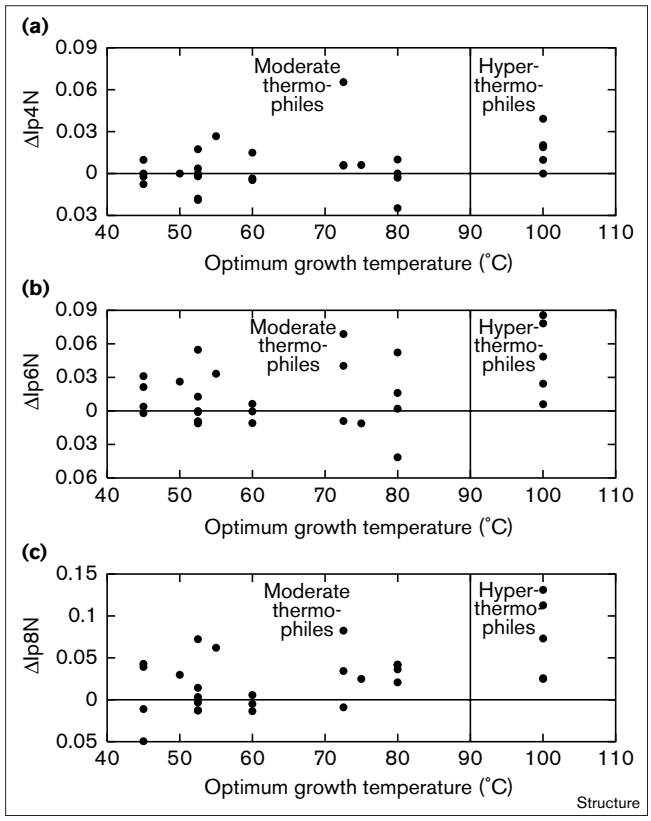
Figure 2 shows the differences between the normalized number of hydrogen bonds (HboN) and unsatisfied hydrogen-bond donors plus acceptors (UhbN) in each thermophilic subunit and the average of its mesophilic homologues as a function of the optimum growth temperature of the source organism. For HboN, the signs of averaged differences are all positive, but all the P values are quite high (i.e., the difference is quite insignificant) as it is also clear from Figure 2. Correlation coefficients between

Figure 2



Differences between **(a)** the normalized number of hydrogen bonds and **(b)** the normalized number of unsatisfied hydrogen-bond donors plus acceptors of each thermophilic subunit and the average of its mesophilic homologues as a function of the optimum growth temperature of the thermophilic source organism.

Figure 3



Differences between the normalized number of ion pairs in each thermophilic subunit and the average of its mesophilic homologues as a function of the optimum growth temperature of the thermophilic source organism. A distance limit of **(a)** 4.0 Å, **(b)** 6.0 Å and **(c)** 8.0 Å was used to define ion pairs.

temperature and differences are also very close to zero. For UhbN, the sign of averaged differences is negative, but the *P* values are relatively high again; correlation coefficients are also negative but not as close to zero as those for HboN. Thus, practically no difference between mesophiles and thermophiles and no tendency to change with temperature is detected in the number of hydrogen bonds; but, for the number of unsatisfied donors and acceptors, there is a slight decrease in thermophiles and a slight tendency to decrease with temperature as well.

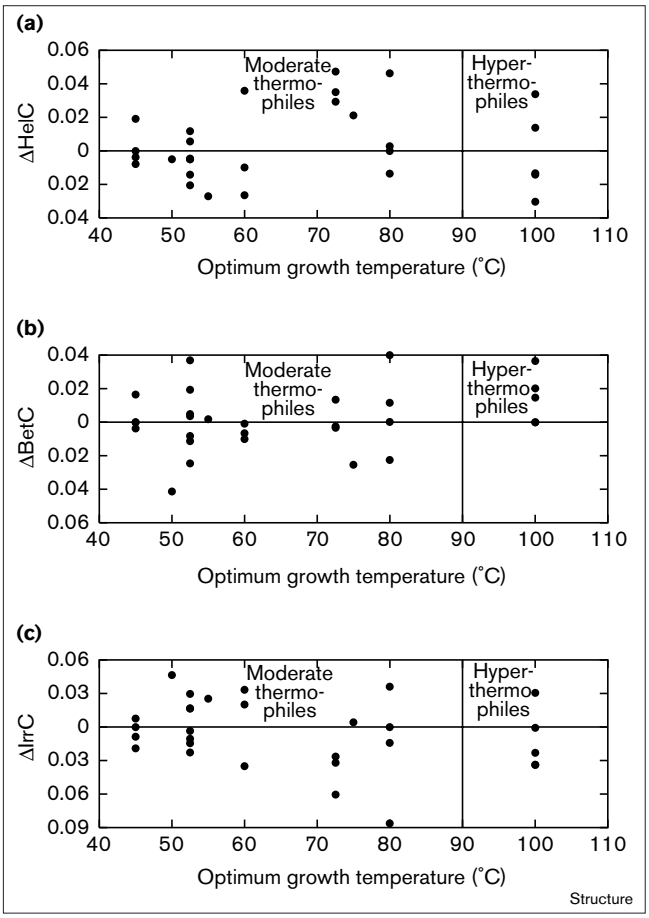
Ion pairs

Figure 3 shows the differences between the normalized number of ion pairs using a distance limit of 4.0, 6.0 and 8.0 Å (Ip4N, Ip6N and Ip8N), respectively, in each thermophilic subunit and the average of its mesophilic homologues as a function of the optimum growth temperature of the source organism. Using a distance limit of 4.0 Å gives only the strongest ion pairs, whereas 6.0 or 8.0 Å also includes weaker ones. As is clear from Figure 3 and the corresponding rows in Table 1, thermophilic proteins definitely tend to contain more ion pairs than mesophilic ones. The differences are very significant, as seen from the extremely low *P* values. With moderately thermophilic proteins ($S_{45–80}$), weaker ion pairs dominate whereas in extremely thermophilic ones (S_{100}) strong ion pairs also have a great role. Besides, the increase in the number of ion pairs correlates relatively well with temperature. On average, the number of ion pairs per residue increases by 0.0003, 0.0006 and 0.0012 for 4.0, 6.0 and 8.0 Å ion pairs (upper distance limit), respectively, with every degree increase in T_{opt}. This means that about four strong and 14 weaker extra ion pairs are expected to appear in a 300-residue protein from an organism with a T_{opt} of about 80°C, compared with its mesophilic homologues with T_{opt} values of about 30°C.

Secondary structure

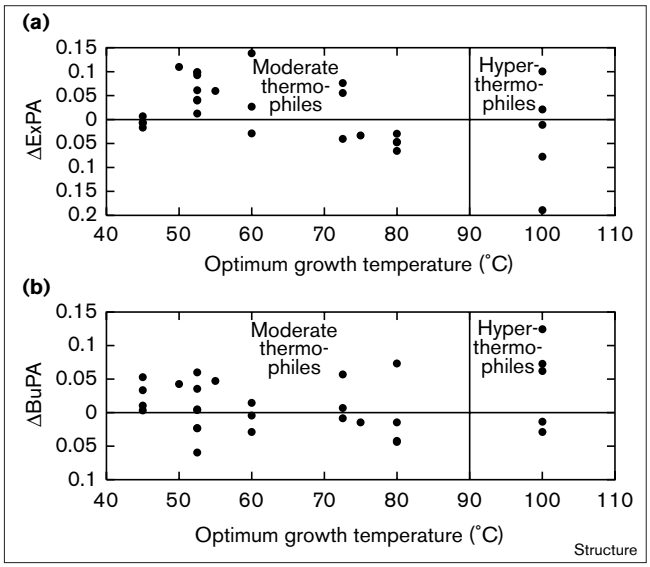
Figure 4 shows the differences between the fraction of helices (HelC), β strands (BetC) and irregular regions (IrrC) of each thermophilic subunit and the average of its mesophilic homologues as a function of the optimum growth temperature of the source organism. See also the corresponding rows in Table 1. As can be seen from the signs of averaged differences, the helix and β content is larger and the fraction of irregular regions is correspondingly smaller in thermophilic proteins than in mesophilic ones. Interestingly, the increase in helix content is greater in moderately thermophilic proteins ($S_{45–80}$), whereas the increase in β content is much more significant in extremely thermophilic ones (S_{100}). It should be noted

Figure 4



Differences between the fraction of **(a)** helical, **(b)** β and **(c)** irregular regions in each thermophilic subunit and the average of its mesophilic homologues as a function of the optimum growth temperature of the thermophilic source organism.

Figure 5



Differences between the polar to apolar surface area ratio calculated for the exposed and buried surface of each thermophilic subunit: **(a)** for ExPA and **(b)** for BuPA. The average of its mesophilic homologues is also shown as a function of the optimum growth temperature of the thermophilic source organism.

that both groups are strongly dominated by α/β type proteins, so this difference cannot be explained by the different secondary structural classes. The *P* values, however, are relatively high, which indicates a high variance. Correlation coefficients show a slight positive correlation between temperature and helix and β content and a slight negative correlation between temperature and fraction of irregular regions.

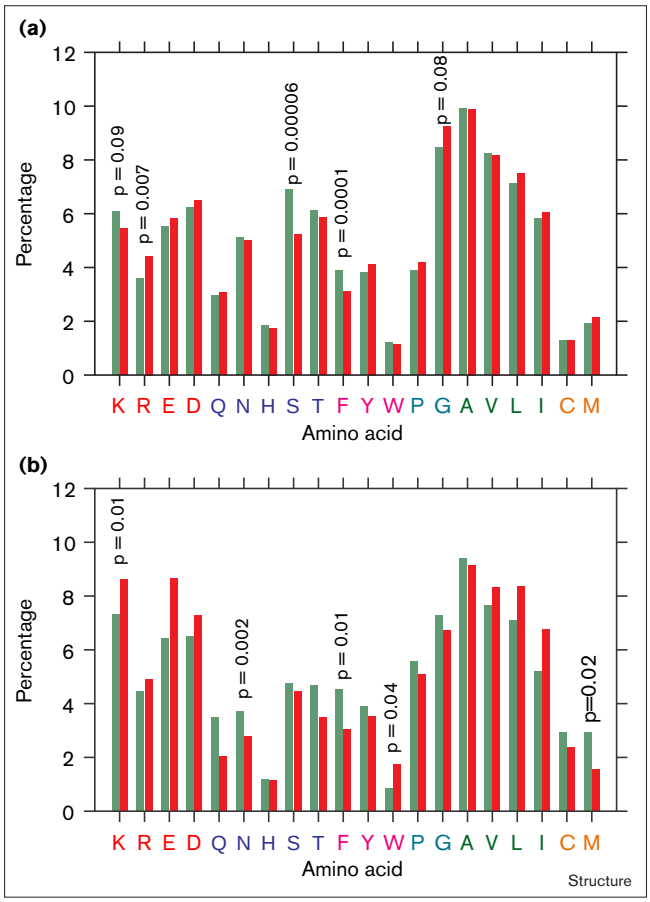
Polar and apolar, exposed and buried surface areas

Figure 5 shows the differences between the polar to apolar surface area ratio for the exposed (ExPA) and buried (BuPA) surface of each thermophilic subunit and the average of its mesophilic homologues as a function of the optimum growth temperature of the source organism (also see Table 1). There is a definite increase in the polarity of the exposed surface in thermophilic subunits, compared with their mesophilic homologues, but the increase is most expressed in moderately thermophilic proteins ($S_{45–80}$). In proteins with T_{opt} above 75°C, an overall polarity increase is not observed and the picture becomes more varied. Because of this, the polarity increase correlates negatively with temperature. In contrast, the buried surface area interestingly becomes more polar in the whole temperature range, although no strong correlation with temperature is observed.

Amino acid composition

Figure 6 shows the amino acid composition of moderately thermophilic ($S_{45–80}$) and extremely thermophilic (S_{100}) proteins and their mesophilic homologues. The *P* values of the most statistically significant ($P < 0.1$) percentage differences are indicated above the corresponding bars in the charts. In general, the percentage of charged residues (lysine, arginine, glutamic acid and asparagine) is higher in thermophilic proteins than in mesophilic ones, but in S_{100} the increase is higher. An interesting exception is lysine, the percentage of which shows a statistically significant decrease in $S_{45–80}$, along with a statistically significant increase in arginine content. In S_{100}, however, the lysine content shows a statistically significant increase whereas the increase in arginine content is smaller and not significant. Phenylalanine content shows a statistically significant decrease in both groups. In $S_{45–80}$, an extremely significant and large decrease in serine content is also observed, which is, however, not present in S_{100}. A significant increase in tryptophan content and a decrease in methionine and asparagine content in S_{100} is also worth mentioning. Expected flexibility altering changes, such as

Figure 6



Comparison of the amino acid compositions of mesophilic (green bars) and thermophilic (red bars) proteins in our data set. Comparison of amino acid compositions of **(a)** moderately thermophilic versus mesophilic proteins and **(b)** hyperthermophilic versus mesophilic proteins. The comparison was performed separately for S_{45-80} and S_{100}. Each bar represents an average of the mean percentages of an amino acid in mesophilic and thermophilic proteins, respectively, over the corresponding subset of protein families. *P* values were calculated from two-tailed, paired *t* tests over the 24 (S_{45-80}) or 5 (S_{100}) pairs of percentages for each amino acid. Significant differences, defined as having a $P < 0.1$, are indicated by the value written over the corresponding pair of bars. Amino acid code letters are shown in different colours according to the physicochemical characteristics of the amino acid (KRED, charged; QNHST, polar; FYW, aromatic; PG, affecting flexibility; AVLI, hydrophobic; CM, sulphur-containing residues).

an increase in proline content and a decrease in glycine content, are only observed in one of the groups each, and they are not significant, except for an unexpected increase in glycine content in S_{45-80}.

Discussion

In this work, we investigated the differences in 13 structural properties between homologous mesophilic and thermophilic proteins. We used high-quality atomic structures for all comparisons.

We observed that proteins from extreme thermophiles ($T_{opt} \approx 100°C$) show characteristically different stabilization patterns in comparison with those from moderately thermophilic organisms ($T_{opt} = 45–80°C$). Although this possibility had been mentioned in the literature before (see [25–30] for studies comparing hyperthermostable proteins with less thermostable ones), we believe that our study is the first to systematically demonstrate this fact. It should be noted, however, that the number of known structures of proteins from organisms with $T_{opt} \approx 100°C$ is very small (there are only five such structures in our data set). This means that our conclusions pertaining to proteins from extremely thermophilic organisms might be of limited validity; in fact, it is possible that much of the observed deviations are because of 'sample bias' (i.e., the peculiarities of the available protein structures).

With this caveat in mind, we can summarize our findings as follows (Table 2). The most significant differences between mesophilic and thermophilic proteins are found in the number of ion pairs, this property also correlates well with growth temperatures. In moderately thermophilic proteins, the number of weaker ion pairs shows significant increase, whereas in extremely thermophilic ones, extra strong ion pairs also appear. A significant decrease is observed in the number of cavities in extremely thermophilic proteins but moderately thermophilic ones do not show any significant differences in cavity parameters in comparison to their mesophilic counterparts. Moderately thermophilic proteins, however, show a significant increase in the polarity of their exposed surface, which is not observed at all with extremely thermophilic ones. A statistically slightly significant increase in all thermophilic proteins in the fraction of α helices and β strands, with a corresponding decrease in irregular regions, is also worth mentioning; with moderately thermophilic proteins, α helical increase is dominant, whereas with extremely thermophilic ones β strand increase is more prevalent. Other properties, such as the number of hydrogen bonds and unsatisfied donors and acceptors as well as the polarity change of buried surfaces, show no or insignificant differences between mesophilic and thermophilic proteins.

Amino acid compositions also exhibit different characteristics in extremely thermophilic proteins in comparison to moderately thermophilic ones. In the latter, the stabilizing effect of lysine→arginine replacements [31] is reflected (lysine content decreases and arginine content increases) whereas in extremely thermophilic proteins, the requirement of a substantial increase in the number of favorable electrostatic interactions appears to be stronger and leads to an increase in the percentages of all charged residues including lysine. A decrease in methionine and asparagine in extremely thermophilic proteins can be explained by the chemical instability of these residues at high temperatures. The serine content shows an apparent decrease in

Table 2

Schematic representation of the findings of this study.

Property		Correlation with temperature	Change in moderately thermophilic proteins	Change in extremely thermophilic proteins
Cavities	Number	↓↓	0	↓↓↓
	Volume	↓	↑	↓
	Area	↓	↑	↓↓
Hydrogen bonds	Number	0	0	0
	Unsatisfied	↓	↓	↓
Ion pairs	<4.0 Å	↑↑	↑	↑↑↑
	<6.0 Å	↑↑	↑↑	↑↑↑
	<8.0 Å	↑↑↑	↑↑↑	↑↑↑
Secondary structure	α	0	↑	0
	β	↑	0	↑↑
	Irregular	↓	↓	↓
Polarity of surfaces	Exposed	↓↓	↑↑↑	0
	Buried	0	↑	↑

Upward arrows refer to positive values and downward arrows refer to negative values; zeros refer to near-zero values. The number of arrows (1, 2 or 3) shows whether the represented value is considered insignificant, moderately significant or highly significant.

moderately thermophilic proteins; in extremely thermophilic ones, this decrease is not observable, probably because the serine content of their mesophilic homologues is already very low.

Comparing our results with those of other similar studies, both agreements and differences are found. Our finding that a decrease in the number and size of internal cavities does not play a significant role in the stabilization of moderately thermophilic proteins confirms the findings of others [22,23]. Our study suggests that extremely thermophilic proteins, however, do utilize this possibility for stabilization.

Ion pairs have long been considered as a possible means to enhance protein thermostability [32]. An increasing body of experimental evidence shows that ion pairs, especially networks of ion pairs, contribute to the increased thermal stability of several thermophilic proteins [6,9,11,26,33] or even play a key role in thermostability [12,13,15,30,34,35]. Systematic studies [20,24] also support this idea. Complex ion pairs [36] are especially likely to enhance stability through a cooperative strengthening mechanism [37]. On the other hand, some theoretical and a number of experimental studies have indicated that salt bridges usually destabilize, or at most slightly stabilize, the native state of proteins [38–41]. In the light of these results, the abundance of ion pairs in thermophilic proteins is somewhat surprising. Most of these studies, however, were made at room temperature.

The major reason for the low stability of salt bridges at room temperature is that the association of two charged residues to form a salt bridge incurs a large desolvation penalty, which is not fully compensated for by favourable interactions within the salt bridge and with the rest of the protein. Theoretical models [42] show that at high temperatures, the desolvation penalty is markedly reduced, and, consequently, salt bridges are preferentially stabilized by high temperatures. Thus, at high temperatures, salt bridges might make a positive contribution to protein stability.

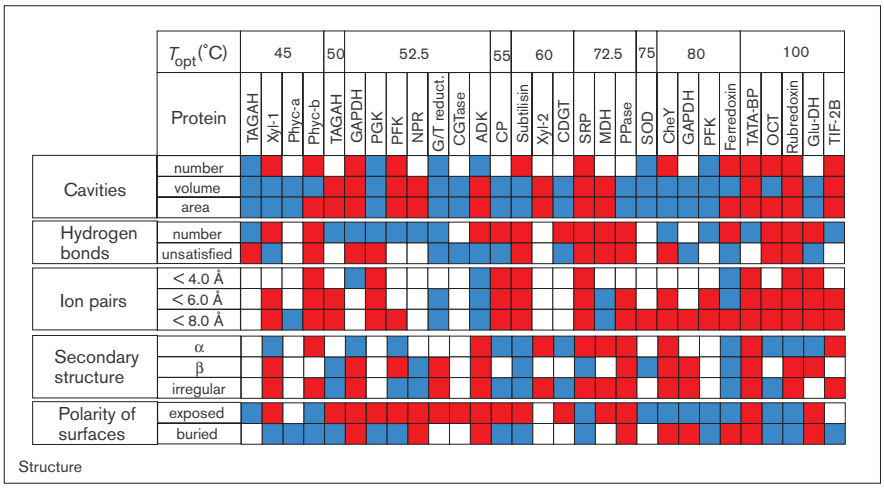
Hydrophobic interactions, rather than ion pairs, could also be expected to increase at higher temperatures. Calculations of Makhatadze and Privalov [43] indicate that the free energy associated with the hydrophobic interaction, which is entropic at room temperature but becomes enthalpic at higher temperatures, reaches its maximum strength around 75°C and starts to weaken at higher temperatures. Because this result comes from a model that contains many assumptions and approximations, the actual temperature of the free energy maximum is debatable and can even vary from protein to protein. Comparisons of mesophilic and thermophilic protein structures [42] indicate that the hydrophobic effect has a higher contribution to stability at higher temperatures. The magnitude of this effect, however, is not large in comparison with the contribution of electrostatic interactions.

According to the model of Elcock [42], there is a significant energetic barrier for breaking a salt bridge, the height of which increases with temperature. A similar barrier is not observed with hydrophobic interactions. This phenomenon also points to the significance of salt bridges in stabilizing proteins at high temperatures.

The results of the present analysis are in accord with the above facts and considerations supporting the key role of electrostatic interactions in hyperthermophilic proteins and they demonstrate that electrostatic interactions are of general importance in both moderately and extremely thermophilic proteins.

The role of hydrogen bonds in the thermal stability of proteins has always been controversial [44,45]. Comparing the results from the studies of the Argos group [23,24] with ours, an interesting discrepancy is found: hydrogen bonds came out in their study as the most important stabilizing factor (they found an increased number of hydrogen bonds in thermophilic proteins in 13 of their 16 protein families), whereas we found practically no statistically significant difference in this parameter between mesophilic and thermophilic proteins. Although we use different methods for evaluation, this fact alone cannot account for the difference in our conclusions. An analysis of the data of Vogt *et al.* [23,24], however, revealed that most of the difference they found between thermophilic and mesophilic proteins in the number of hydrogen bonds is caused by using bad-quality, erroneous protein structures for the comparisons. These bad-quality or incomplete

Figure 7



Colour-coded diagram demonstrating the variety of 'stabilization strategies' utilized by thermophilic protein subunits. Each square in the table is coloured according to how the given property (as shown at the left side of the row) is changed in a given thermophilic subunit (as shown at the top of the column, see the Materials and methods section for the abbreviations) in comparison to the average of its mesophilic homologues. A red square indicates a change in stabilizing direction, a blue square indicates a change in destabilizing direction and a white square means that there is no significant difference. (With cavity parameters, number of unsatisfied hydrogen-bond donors plus acceptors, fraction of irregular structure and polarity of buried surface, an increase is considered destabilizing and a decrease stabilizing; with the remaining properties, the reverse is true.)

structures include PDB entries 4gpd, 3gpd, 1llc, 5ldh, 1fdx, 3sdp and 3pgk; these are all mesophilic proteins that contain markedly (often strikingly) fewer hydrogen bonds than their mesophilic and thermophilic homologues. This is clearly because of their bad quality, and, in some cases, too many missing atoms. On the other hand, the data set of Vogt *et al.* [23,24] contains only a single bad-quality thermophilic structure (1ril), which leads to a serious imbalance and leads the authors to conclude that mesophilic proteins contain fewer hydrogen bonds than thermophilic ones. But this is an artifact; if these bad-quality structures had been omitted from the database of Vogt *et al.* [23,24], then the observed difference in the number of hydrogen bonds would have become insignificant and the authors would have arrived at the same conclusions as ourselves. We think that our conclusion about the hydrogen bonds is correct; we excluded all bad-quality and incomplete structures when creating our data set.

Although other parameters might not be as sensitive to quality as hydrogen bonds are, bad-quality structural data obviously are a source of errors. Therefore, strict quality control is important in studies like this. We suggest using quality-checking software such as WHAT_CHECK and preferably the PDBREPORT database [46] for studies that analyze structural parameters on large sets of proteins.

Although the only property showing a relatively strong and clear-cut correlation with growth temperature is the number of ion pairs, and other properties show much smaller statistical significances in our analysis, this fact does not justify the conclusion that changes in these other properties have no stabilizing contributions. On the contrary, most of them do have such contributions, as is revealed by site-directed mutagenesis experiments (see [47] for a review). What is found here is that most properties are utilized for stabilization only in some protein families, and when the differences in a property are evaluated for the whole unified data set then the stabilizing changes found in these families are compensated for by opposite changes in other families; consequently, in the final analysis, no statistically significant difference is found. At present, individual protein families do not contain enough known thermophilic and mesophilic structures to perform a similar statistical analysis for the individual families, which certainly would reveal the individual 'stabilization strategy' used by each family (if there exists a family-wide strategy at all). Such statistical analyses will, however, be feasible in the future when more atomic structures in individual protein families become available.

At present, however, we can rely on what we know from experiments about the physical interactions determining protein stability and can assume that cavities and buried polar residues usually destabilize whereas hydrogen bonds and secondary structural elements generally have a stabilizing contribution. Using these considerations, in Figure 7 we tried to visualize the variety of 'stabilization strategies' or patterns observed with our set of thermophilic protein subunits, using different colours for stabilizing and destabilizing changes. The diagram shows that almost every protein family uses an individual 'strategy' to achieve a high thermal stability. Entirely different strategies can be observed. For example, ion pairs give a positive contribution to the extra stability of pyrophospatase from *Thermus thermophilus* (relative to the pyrophosphatase from *Escherichia coli*) whereas it is somewhat destabilized by having larger cavities. For malate dehydrogenase from *Thermus flavus*, however, the reverse occurs: it has fewer ion pairs than its mesophilic

counterpart (from pig) and its cavities are smaller. Additional examples can be seen in the diagram.

Biological implications

One of the most important means of evolutionary adaptation to high temperatures (thermophily) is the enhancement via mutations of the intrinsic thermal stability of proteins [1]. In this study, comparative analyses of a number of molecular properties were performed on a data set comprising all available high-quality thermophilic protein structures and their mesophilic counterparts, in order to obtain a better understanding of the mechanisms of thermophilic adaptation of proteins.

Despite the great variability of stabilization 'strategies' of the individual proteins (Figure 7), some general trends can be seen (Table 2). The strongest correlation between thermostability and a structural parameter is observed in the case of electrostatic interactions, but other factors, such as secondary structure, cavities, and polar surface fraction, also show some correlation with thermostability. Moderately and extremely thermostable proteins appear to rely on somewhat different mechanisms to achieve greater stability. The reason for this might lie in the temperature dependence of the various forces involved in protein stabilization, but it might also reflect the fact that extremely thermophilic organisms all belong to the domain archaea, and are therefore phylogenetically distinct from moderately thermophilic organisms, which are non-archaea. It should be noted that although bacterial and archaeal hyperthermophiles are close to the root of the phylogenetic tree, preceding their mesophilic counterparts [48], recent analyses still suggest that extant hyperthermophiles evolved from mesophiles via adaptation to high temperature [49]. From the observed patterns of stabilization, it is reasonable to assume that those interactions are primarily utilized for enhancing thermostability that can form relatively rapidly in the process of molecular evolution. Therefore, interactions forming on the outer surface of a protein molecule are preferred. The knowledge of the evolutionary mechanisms for protein thermostabilization, some of which are revealed by this study, can help us to engineer proteins with enhanced thermostability for practical purposes.

Materials and methods

Construction of a data set

A data set containing homologous mesophilic and thermophilic protein structures was constructed by the following procedure. The names of source organisms of proteins in the November 1998 version of the PDB were extracted. Mesophilic organisms were deleted from the list and only thermophilic organisms, defined as having an optimal growth temperature higher than 45°C, were retained. For microorganisms, optimal growth temperatures were determined from the literature [50,51]; if several different and equally reliable values or ranges were found, then the average value was taken. The number of thermophilic organisms was 36. Structures in the PDB from these source organisms were retrieved (209 structures). The list was further reduced by eliminating multiple structures of the same protein, retaining the one with the highest resolution (when it was a mutant, however, then the wild-type structure was chosen instead). The resulting 103 structures (173 subunits) were divided into structural families using the FSSP database [52]. FSSP entries containing the thermophilic proteins also contain all their structural homologues by definition. After filtering out identical subunits (only one subunit was retained from subunits having identical amino acid sequences, point mutations ignored), the resulting data set contained 111 protein subunits from 94 protein families.

In the next step, the 94 FSSP files were examined and all the families that did not contain any mesophilic structures homologous to the thermophilic ones were excluded from the data set. Two structures within an FSSP file were considered homologous if their size was about the same and their sequence identity was at least 30%. (FSSP files often contain proteins that are partially structurally similar to each other but their size or their overall fold is essentially different and the sequence identity between them is very low. The 30% identity limit effectively selects proteins that have the same basic fold.) A number of families (50) were discarded because they did not contain suitable mesophilic structures. Two more families were discarded because their mesophilic and thermophilic structures have been determined at very different temperatures. The resulting data set contained 42 families. In the mesophilic subset of the data set, multiple structures and identical subunits were eliminated the same way as done earlier for the thermophilic proteins.

Structures were checked for missing atoms and chain breaks. Any structures with chain breaks or more than three incomplete sidechains were excluded from the data set; in these cases, they were replaced by another structure of the same protein if such a structure was available. In this procedure, eight families had to be completely excluded from the data set. In a few cases, less than three incomplete sidechains were present; the missing atoms were then generated using the 'replace residue' command of the program InsightII (MSI, Inc.). In the case of PDB entries 1pca and 1nsa, which are proteases with their prosegments, the N-terminal prosegments were removed so that the structures could fit with the remaining structures (having no prosegments) in the family.

The quality of the proteins in the remaining 34 families was checked using the WHAT_CHECK program [46]; for most structures, the quality reports in the PDBREPORT database [46] could be used. All structures that had the qualification 'bad' in the quality report, or had a resolution of 3.0 Å or worse, were excluded from the data set (as always, they were replaced by other structures of the same protein if this was possible). In the end, nine families had to be discarded this way and the final version of the data set contained 25 families.

The final data set

The final data set, containing 64 mesophilic and 29 thermophilic subunit structures in 25 protein families, is given below. For each family, the family name and abbreviation (used in Figure 7) are shown in bold type and the PDB identifiers with subunit identifiers (where applicable) of the structures follow. Thermophilic structures are indicated in italic type and the optimum growth temperature (in °C) of the source organism is given for each one.

1. Transcription initation factor IIb (TIF-2B): 1volA, *1aisB*/100; 2. Superoxide dismutase (SOD): 1abmA, 1ar4A, 1idsA, 1isaA, 1vewA, *3mdsA*/75; 3. Glutamate dehydrogenase (Glu-DH): 1hrdA, *1gtmA*/100; 4. Malate dehydrogenase (MDH): 4mdhA, *1bmdA*/72.5; 5. Phycocyanin alpha chain (Phyc-a): 1cpcA, 1liaA, 1allA, *1phnA*/45; 6. Signal recognition particle (receptor) (SRP): 1fts, *1ffh*/72.5; 7. Ferredoxin: 1fxd, 1fxrA, *1vjw*/80; 8. Subtilisin: 1sup, 1cseE, 1bh6, 1svn, 2pkc, 1sbnE, 1meeA, *1thm*/60; 9. Neutral protease (thermolysin) (NPR): 1npc, *1lnfE*/52.5; 10. Rubredoxin: 1iro, 1rdg, 6rxn, 8rxnA, *1caa*/100; 11. Cyclodextrin glycosyltransferase (CGTase): 1cdg, 1cgt, 1pamA, *1ciu*/60, *1cyg*/52.5; 12. Phycocyanin beta chain (Phyc-b):

1allB, 1cpcB, 1liaB, *1phnB*/45; 13. 3-Phosphoglycerate kinase (PGK): 1qpg, *1php*/52.5, *1vpe*/80; 14. Glyceraldehyde-3-phosphate dehydrogenase (GAPDH): 1a7kA, 1gadO, 1szjG, *1gd1O*/52.5, *1hdgO*/80; 15. Xylanase (I) (Xyl-1): 1enxA, 1ukrA, 1xnb, 1xnd, 1xyn, *1yna*/45; 16. Xylanase (II) (Xyl-2): 1clxA, 2exo, *1xyzA*/60; 17. TATA box binding protein (TATA-BP): 1cdwA, 1vokA, *1pczA*/100; 18. Adenylate kinase (ADK): 1ak2, 2ak3A, 1aky, 1ukz, 1akeA, 3ukd, *1zip*/52.5; 19. Carboxypeptidase (CP): 2ctc, 1nsa, 1pca, *1obr*/55; 20. Ornithine carbamoyltransferase (OCT): 2otcA, *1a1s*/100; 21. Pyrophosphatase (PPase): 1obwA, *2prd*/72.5; 22. CheY protein (CheY): 3chy, 2chf, *1tmy*/80; 23. Glutathione/trypanothione reductase (G/T reduct.): 1aogA, 1febA, 1gerA, 3grs, *1ebdA*/52.5; 24. Phosphofructokinase (PFK): 1pfkA, *4pfk*/52.5; 25. Triacylglycerol acylhydrolase (TAGAH): 1lgyA, *1tib*/50, *3tgl*/45.

Cavities

The number, surface areas and volumes of cavities were calculated by the Molecular Surface Package [53]. The molecular surface area was used instead of the accessible surface area because the former was shown to correlate better with thermodynamic quantities than the latter [54]. A probe radius of 1.2 Å was used throughout because this radius was shown to be optimal in defining cavities; this value provides the best conservation of cavities among homologous proteins [55].

Hydrogen bonds

The number of hydrogen bonds and unsatisfied hydrogen-bond donors and acceptors were determined using the HB2 algorithm of the WHAT IF molecular modeling package [56]. This program uses a special forcefield and a sophisticated algorithm to find the best hydrogen-bond network and is believed to be more reliable than the simple distance-criterion based definition of hydrogen bonds.

Ion pairs

Ion pairs were defined using a simple distance criterion: two oppositely charged residues were considered an ion pair if their closest oppositely charged atoms were closer to each other than a predefined limit distance. The distance limit was chosen to be 4.0 Å (the usual value, see [57]) to find stronger ion pairs; to allow for weaker electrostatic interactions, distance limits of 6.0 Å and 8.0 Å were also used. Ion pairs created by histidine were not considered, because it is problematic to decide whether a given histidine residue is charged in a protein.

Polar and apolar, exposed and buried surface areas

Surface areas were calculated by the WHAT IF program [56] using the highest available precision and a probe radius of 1.4 Å (the standard water radius usually used for surface calculations). N and O atoms were considered polar, the rest apolar. To calculate the buried surfaces, the model of the polypeptide chain was unfolded by setting all torsion angles to the values given in [58] and the surfaces for this extended chain were calculated. Buried surfaces were obtained by subtracting the values calculated for the folded chain from those calculated for the unfolded chain. This procedure was shown to give values that correlate better with thermodynamic quantities than the usual procedure that uses a glycine–X–glycine tripeptide as the model of the unfolded state of each residue [58,59].

As we indicated earlier, this study only considers isolated subunits. In the case of oligomeric proteins, therefore, surfaces for a subunit were calculated in the absence of other subunits. Although polar to apolar surface area ratios are expected to differ between monomeric proteins and subunits of oligomeric proteins, this does not affect our comparisons because mesophilic and thermophilic proteins are only compared within structural families (i.e., groups characterized by the same quaternary structure). Besides, subunits of oligomeric proteins are expected to be capable of assuming a native-like fold on their own before association, that is, they can utilize the same means for stabilization as monomeric proteins do.

Secondary structure

The secondary structures were calculated by the DSSP program [60]. Residues having a letter H or G in the DSSP output were considered to be in helices; those having E or B were considered to be in β structures. All remaining residues were considered to be in irregular regions. The percentage (fraction of chain length) of each secondary structural element was determined.

Evaluation of data

The following quantities were calculated for all protein subunits in the data set: number, total surface area and total volume of internal cavities; the number of hydrogen bonds and unsatisfied hydrogen bond donors plus acceptors; the number of ion pairs using a limit distance of 4.0, 6.0 and 8.0 Å for definition; buried and exposed polar to apolar surface area ratio; fraction of helical, β and irregular regions; amino acid composition (i.e., fraction of each residue type relative to chain length). Quantities that increase with protein size (i.e., number, total volume and total surface area of cavities, number of hydrogen bonds, number of unsatisfied donors plus acceptors, and number of ion pairs) were normalized by the averaged number of residues of all proteins in each protein family so that quantities calculated for proteins in different families could be compared. Although it would have been more straightforward to normalize each parameter by the chain length of each individual protein subunit, this could have led to artifacts because thermophilic chains might tend to be shorter than their mesophilic homologues (because of shorter loops [61]), and this would artificially increase the normalized parameters for the thermophilic subunits relative to the mesophilic ones. This was the reason we chose the normalization method described above, that is, normalize by the average chain length of each family.

Our data set contains 29 thermophilic structures in the 25 protein families; 21 families only contain one thermophilic structure and the remaining four families contain two thermophilic structures with different optimum growth temperatures (T_{opt}) of the source organisms. To be able to recognize correlations between the calculated properties and T_{opt}, the two thermophilic structures in each of the mentioned four families were considered separately instead of averaging their properties. For each thermophilic protein subunit and each property, the property value calculated for the thermophilic protein subunit and the average value of the property calculated for the mesophilic protein subunits in the same family was computed. This gave 29 pairs of numbers for each property. The differences between the numbers in each pair were also calculated and shown as graphs (Figures 1–5) and the correlation coefficients between T_{opt} and these differences were calculated. A two-tailed, paired *t* test of the 29 pairs of numbers was used to evaluate the statistical significance of the difference between the thermophilic and mesophilic proteins for the entire data set. This test, and the calculation of correlation coefficients mentioned above were also performed for the two subsets of the data set, $S_{45–80}$ and S_{100} (moderately and extremely thermophilic proteins), defined in the Results section.

Supplementary material

Supplementary material including a table describing in detail the data set used in our analysis is available at http://current-biology.com/supmat/supmatin.htm.

Acknowledgements

We thank Ferenc Vonderviszt (University of Veszprém), David Eisenberg, Richard E Dickerson and Cameron Mura (UCLA-DOE Laboratory of Structural Biology and Molecular Medicine) for their suggestions made after carefully reading of an earlier version of the manuscript. This work was funded by grants OTKA F 020874 and T 022370 as well as FKFP 0166/97. AS was supported by a Magyary Zoltán postdoctoral fellowship.

References

1. Jaenicke, R. (1991). Protein stability and molecular adaptation to extreme conditions. *Eur. J. Biochem.* **202**, 715-728.
2. Jaenicke, R. & Böhm, G. (1998). The stability of proteins in extreme environments. *Curr. Opin. Struct. Biol.* **8**, 738-748.

3. Jaenicke, R. & Závodszky, P. (1990). Proteins under extreme physical conditions. *FEBS Lett.* **268**, 344-349.
4. Russell, R.J.M. & Taylor, G.L. (1995). Engineering thermostability: lessons from thermophilic proteins. *Curr. Opin. Biotechnol.* **6**, 370-374.
5. Querol, E., Perez-Pons, J.A. & Mozo-Villarias, A. (1996). Analysis of protein conformational characteristics related to thermostability. *Protein Eng.* **9**, 265-271.
6. Walker, J.E., Wonacott, A.J. & Harris, J.I. (1980). Heat stability of a tetrameric enzyme, D-glyceraldehyde-3-phosphate dehydrogenase. *Eur. J. Biochem.* **108**, 581-586.
7. Davies, G.J., Gamblin, S.J., Littlechild, J.A. & Watson, H.C. (1993). The structure of a thermally stable 3-phosphoglycerate kinase and a comparison with its mesophilic equivalent. *Proteins* **15**, 283-289.
8. Fujinaga, M., Berthet-Colominas, C., Yaremchuk, A.D., Tukalo, M.A. & Cusack, S. (1993). Refined crystal structure of the seryl-tRNA synthetase from *Thermus thermophilus* at 2.5 Å resolution. *J. Mol. Biol.* **234**, 222-233.
9. Chan, M.K., Mukund, S., Kletzin, A., Adams, M.W. & Rees, D.C. (1995). Structure of a hyperthermophilic tungstopterin enzyme, aldehyde ferredoxin oxidoreductase. *Science* **267**, 1463-1469.
10. Hecht, H.J., Erdmann, H., Park, H.J., Sprinzl, M. & Schmid, R.D. (1995). Crystal structure of NADH oxidase from *Thermus thermophilus*. *Nat. Struct. Biol.* **2**, 1109-1114.
11. Korndörfer, I., Steipe, B., Huber, R., Tomschy, A. & Jaenicke, R. (1995). The crystal structure of holo-glyceraldehyde-3-phosphate dehydrogenase from the hyperthermophilic bacterium *Thermotoga maritima* at 2.5 Å resolution. *J. Mol. Biol.* **246**, 511-521.
12. Yip, K.S., *et al.*, & Consalvi, V. (1995). The structure of *Pyrococcus furiosus* glutamate dehydrogenase reveals a key role for ion-pair networks in maintaining enzyme stability at extreme temperatures. *Structure* **3**, 1147-1158.
13. Aguilar, C.F., *et al.*, & Pearl, L.H. (1997). Crystal structure of the β-glycosidase from the hyperthermophilic archeon *Sulfolobus solfataricus*: resilience as a key factor in thermostability. *J. Mol. Biol.* **271**, 789-802.
14. Harris, G.W., *et al.*, & Perez, S. (1997). Structural basis of the properties of an industrially relevant thermophilic xylanase. *Proteins* **29**, 77-86.
15. Russell, R.J., Ferguson, J.M., Hough, D.W., Danson, M.J. & Taylor, G.L. (1997). The crystal structure of citrate synthase from the hyperthermophilic archaeon *Pyrococcus furiosus* at 1.9 Å resolution. *Biochemistry* **36**, 9983-9994.
16. Wallon, G., Kryger, G., Lovett, S.T., Oshima, T., Ringe, D. & Petsko, G.A. (1997). Crystal structures of *Escherichia coli* and *Salmonella typhimurium* 3-isopropylmalate dehydrogenase and comparison with their thermophilic counterpart from *Thermus thermophilus*. *J. Mol. Biol.* **266**, 1016-1031.
17. Argos, P., Rossmann, M.G., Grau, U.M., Zuber, H., Frank, G. & Tratschin, J.D. (1979). Thermal stability and protein structure. *Biochemistry* **18**, 5698-5703.
18. Menendez-Arias, L. & Argos, P. (1989). Engineering protein thermal stability. Sequence statistics point to residue substitutions in α-helices. *J. Mol. Biol.* **206**, 397-406.
19. Böhm, G. & Jaenicke, R. (1994). Relevance of sequence statistics for the properties of extremophilic proteins. *Int. J. Pept. Protein Res.* **43**, 97-106.
20. Spassov, V.Z., Karshikoff, A.D. & Ladenstein, R. (1995). The optimization of protein–solvent interactions: thermostability and the role of hydrophobic and electrostatic interactions. *Protein Sci.* **4**, 1516-1527.
21. Warren, G.L. & Petsko, G.A. (1995). Composition analysis of alpha-helices in thermophilic organisms. *Protein Eng.* **8**, 905-913.
22. Karshikoff, A. & Ladenstein, R. (1998). Proteins from thermophilic and mesophilic organisms essentially do not differ in packing. *Protein Eng.* **11**, 867-872.
23. Vogt, G. & Argos, P. (1997). Protein thermal stability: hydrogen bonds or internal packing? *Fold. Des.* **2**, S40-S46.
24. Vogt, G., Woell, S. & Argos, P. (1997). Protein thermal stability, hydrogen bonds, and ion pairs. *J. Mol. Biol.* **269**, 631-643.
25. Blamey, J.M., Mukund, S. & Adams, M.W. (1994). Properties of a thermostable 4Fe-ferredoxin from the hyperthermophilic bacterium *Thermotoga maritima*. *FEMS Microbiol. Lett.* **121**, 165-169.
26. Tomschy, A., Böhm, G. & Jaenicke, R. (1994). The effect of ion pairs on the thermal stability of D-glyceraldehyde 3–phosphate dehydrogenase from the hyperthermophilic bacterium *Thermotoga maritima*. *Protein Eng.* **7**, 1471-1478.
27. Kletzin, A. & Adams, M.W. (1996). Molecular and phylogenetic characterization of pyruvate and 2–ketoisovalerate ferredoxin oxidoreductases from *Pyrococcus furiosus* and pyruvate ferredoxin oxidoreductase from *Thermotoga maritima*. *J. Bacteriol.* **178**, 248-257.
28. Auerbach, G., *et al.*, & Jacob, U. (1997). Closed structure of phosphoglycerate kinase from *Thermotoga maritima* reveals the catalytic mechanism and determinants of thermal stability. *Structure* **5**, 1475-1483.
29. Auerbach, G., *et al.*, & Jaenicke, R. (1998). Lactate dehydrogenase from the hyperthermophilic bacterium *Thermotoga maritima*: the crystal structure at 2.1 Å resolution reveals strategies for intrinsic protein stabilization. *Structure* **6**, 769-781.
30. Vetriani, C., *et al.*, & Robb, F.T. (1998). Protein thermostability above 100°C: a key role for ionic interactions. *Proc. Natl Acad. Sci. USA* **95**, 12300-12305.
31. Mrabet, N.T., *et al.*, & Wodak, S.J. (1992). Arginine residues as stabilizing elements in proteins. *Biochemistry* **31**, 2239-2253.
32. Perutz, M.F. & Raidt, H. (1975). Stereochemical basis of heat stability in bacterial ferredoxins and in haemoglobin A2. *Nature* **255**, 256-259.
33. Kelly, C.A., Nishiyama, M., Ohnishi, Y., Beppu, T. & Birktoft, J.J. (1993). Determinants of protein thermostability observed in the 1.9 Å crystal structure of malate dehydrogenase from the thermophilic bacterium *Thermus flavus*. *Biochemistry* **32**, 3913-3922.
34. Goldman, A. (1995). How to make my blood boil. *Structure* **3**, 1277-1279.
35. Rice, D.W., *et al.*, & Engel, P.C. (1996). Insights into the molecular basis of thermal stability from the structure determination of *Pyrococcus furiosus* glutamate dehydrogenase. *FEMS Microbiol. Rev.* **18**, 105-117.
36. Musafia, B., Buchner, V. & Arad, D. (1995). Complex salt bridges in proteins: statistical analysis of structure and function. *J. Mol. Biol.* **254**, 761-770.
37. Horovitz, A., Serrano, L., Avron, B., Bycroft, M. & Fersht, A.R. (1990). Strength and co-operativity of contributions of surface salt bridges to protein stability. *J. Mol. Biol.* **216**, 1031-1044.
38. Sali, D., Bycroft, M. & Fersht, A.R. (1991). Surface electrostatic interactions contribute little of stability of barnase. *J. Mol. Biol.* **220**, 779-788.
39. Sun, D.P., Sauer, U., Nicholson, H. & Matthews, B.W. (1991). Contributions of engineered surface salt bridges to the stability of T4 lysozyme determined by directed mutagenesis. *Biochemistry* **30**, 7142-7153.
40. Hendsch, Z.S. & Tidor, B. (1994). Do salt bridges stabilize proteins? A continuum electrostatic analysis. *Protein Sci.* **3**, 211-226.
41. Waldburger, C.D., Schildbach, J.F. & Sauer, R.T. (1995). Are buried salt bridges important for protein stability and conformational specificity? *Nat. Struct. Biol.* **2**, 122-128.
42. Elcock, A.H. (1998). The stability of salt bridges at high temperatures: implications for hyperthermophilic proteins. *J. Mol. Biol.* **284**, 489-502.
43. Makhatadze, G.I. & Privalov, P.L. (1995). Energetics of protein structure. *Adv. Protein Chem.* **47**, 307-425.
44. Dill, K.A. (1990). Dominant forces in protein folding. *Biochemistry* **29**, 7133-7155.
45. Pace, C.N., Shirley, B.A., McNutt, M. & Gajiwala, K. (1996). Forces contributing to the conformational stability of proteins. *FASEB J.* **10**, 75-83.
46. Hooft, R.W., Vriend, G., Sander, C. & Abola, E.E. (1996). Errors in protein structures. *Nature* **381**, 272.
47. Fersht, A.R. & Serrano, L. (1993). Principals of protein stability derived from protein engineering experiments. *Curr. Opin. Struct. Biol.* **3**, 75-83.
48. Woese, C. (1998). The universal ancestor. *Proc. Natl Acad. Sci. USA* **95**, 6854-6859.
49. Galtier, N., Tourasse, N. & Gouy, M. (1999). A nonhyperthermophilic common ancestor to extant life forms. *Science* **283**, 220-221.
50. Holt, J.G. (1989). *Bergey's Manual of Systematic Bacteriology*. Third Edition, Williams & Wilkins, Baltimore.
51. Holt, J.G. (1994). *Bergey's Manual of Determinative Bacteriology*. Ninth edition, Williams & Wilkins, Baltimore.
52. Holm, L., Ouzounis, C., Sander, C., Tuparev, G. & Vriend, G. (1992). A database of protein structure families with common folding motifs. *Protein Sci.* **1**, 1691-1698.
53. Connolly, M.L. (1993). The molecular surface package. *J. Mol. Graph.* **11**, 139-141.

54. Tunon, I., Silla, E. & Pascual-Ahuir, J.L. (1992). Molecular surface area and hydrophobic effect. *Protein Eng.* **5**, 715-716.
55. Hubbard, S.J., Gross, K.H. & Argos, P. (1994). Intramolecular cavities in globular proteins. *Protein Eng.* **7**, 613-626.
56. Vriend, G. (1990). WHAT IF: a molecular modeling and drug design program. *J. Mol. Graph.* **8**, 52-56.
57. Barlow, D.J. & Thornton, J.M. (1983). Ion-pairs in proteins. *J. Mol. Biol.* **168**, 867-885.
58. Oobatake, M. & Ooi, T. (1993). Hydration and heat stability effects on protein unfolding. *Prog. Biophys. Mol. Biol.* **59**, 237-284.
59. Creamer, T.P., Srinivasan, R. & Rose, G.D. (1995). Modeling unfolded states of peptides and proteins. *Biochemistry* **34**, 16245-16250.
60. Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577-2637.
61. Thompson, M.J. & Eisenberg, D. (1999). Transproteomic evidence of a loop-deletion mechanism for enhancing protein thermostability. *J. Mol. Biol.* **290**, 595-604.

Because *Structure with Folding & Design* operates a 'Continuous Publication System' for Research Papers, this paper has been published on the internet before being printed (accessed from http://biomednet.com/cbiology/str). For further information, see the explanation on the contents page.

Supplementary material

Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits: results of a comprehensive survey

András Szilágyi and Péter Závodszky

Structure 2000, **8**:493–504

Table S1

The data set used for this study. The optimum growth temperatures of thermophilic organisms are given; mesophilic organisms are indicated as 'meso'.

Family name	PDB ID, source organism, T_{opt}, resolution
1. Transcription initation factor IIb (TIF-2B)	1volA (Human, meso) 2.7 Å 1aisB (*Pyrococcus woesei*, 100 °C) 2.1 Å
2. Superoxide dismutase (Mn- or Fe-dependent) (SOD)	1abmA (Human, meso) 2.2 Å 1ar4A (*Propionibacterium freudenreichii*, meso) 1.9 Å 1idsA (*Mycobacterium tuberculosis*, meso) 2.0 Å 1isaA (*Escherichia coli*, meso) 1.8 Å 1vewA (*Escherichia coli*, meso) 2.1 Å 3mdsA (*Thermus thermophilus*, 75°C) 1.8 Å
3. Glutamate dehydrogenase (Glu-DH)	1hrdA (*Clostridium symbiosum*, meso) 1.96 Å 1gtmA (*Pyrococcus furiosus*, 100°C) 2.2 Å
4. Malate dehydrogenase (MDH)	4mdhA (Pig heart, meso) 2.5 Å 1bmdA (*Thermus flavus*, 72.5°C) 1.9 Å
5. Phycocyanin alpha chain (Phyc-a)	1cpcA (*Fremyella diplosiphon*, meso) 1.66 Å 1liaA (*Polysiphonia urceolata*, meso) 2.8 Å 1allA (*Spirulina platensis*, meso) 2.3 Å 1phnA (*Cyanidium caldarium*, 45°C) 1.65 Å
6. Signal recognition particle (receptor) (SRP)	1fts (*Escherichia coli*, meso) 2.2 Å 1ffh (*Thermus aquaticus*, 72.5°C) 2.05 Å
7. Ferredoxin	1fxd (*Desulfovibrio gigas*, meso) 1.7 Å 1fxrA (*Desulfovibrio africanus*, meso) 2.3 Å 1vjw (*Thermotoga maritima*, 80°C) 1.75 Å
8. Subtilisin	1sup (*Bacillus amyloliquefaciens*, meso) 1.6 Å 1cseE (*Bacillus subtilis*, meso) 1.2 Å 1bh6 (*Bacillus licheniformus*, meso) 1.75 Å 1svn (*Bacillus lentus*, meso) 1.4 Å 2pkc (*Tritirachium album limber*, meso) 1.5 Å 1sbnE (*Bacillus subtilis*, meso) 2.1 Å 1meeA (*Bacillus mesentericus*, meso) 2.0 Å 1thm (*Thermoactinomyces vulgaris*, 60°C) 1.37Å
9. Neutral protease (thermolysin) (NPR)	1npc (*Bacillus cereus*, meso) 2.0 Å 1lnfE (*Bacillus thermoproteolyticus*, 52.5°C) 1.7 Å
10. Rubredoxin	1iro (*Clostridium pasteurianum*, meso) 1.1 Å 1rdg (*Desulfovibrio gigas*, meso) 1.4 Å 6rxn (*Desulfovibrio desulfuricans*, meso) 1.5 Å 8rxnA (*Desulfovibrio vulgaris*, meso) 1.0 Å 1caa (*Pyrococcus furiosus*, 100°C) 1.8 Å
11. Cyclodextrin glycosyltransferase (CGTase)	1cdg (*Bacillus circulans* strain 251, meso) 2.0 Å 1cgt (*Bacillus circulans* strain 8, meso) 2.0 Å 1pamA (*Bacillus sp.* 1011, meso) 1.8 Å 1ciu (*Thermoanaerobacterium thermosulfurigenes*, 60°C) 2.3 Å 1cyg (*Bacillus stearothermophilus*, 52.5°C) 2.5 Å
12. Phycocyanin beta chain (Phyc-b)	1allB (*Spirulina platensis*, meso) 2.3 Å 1cpcB (*Fremyella diplosiphon*, meso) 1.66 Å 1liaB (*Polysiphonia urceolata*, meso) 2.8 Å 1phnB (*Cyanidium caldarium*, 45°C) 1.65 Å

Table S1 (continued)

Family name	PDB ID, source organism, T_{opt}, resolution
13. 3-Phosphoglycerate kinase (PGK)	1qpg (Yeast, meso) 2.4 Å 1php (*Bacillus stearothermophilus*, 52.5°C) 1.65 Å 1vpe (*Thermotoga maritima*, 80°C) 2.0 Å
14. Glyceraldehyde-3-phosphate dehydrogenase (GAPDH)	1a7kA (*Leishmania mexicana*, meso) 2.8 Å 1gadO (*Escherichia coli*, meso) 1.8 Å 1szjG (*Palinurus versicolor*, meso) 2.0 Å 1gd1O (*Bacillus stearothermophilus*, 52.5°C) 1.8 Å 1hdgO (*Thermotoga maritima*, 80°C) 2.5 Å
15. Xylanase (I) (Xyl-1)	1enxA (*Trichoderma reesei*, meso) 1.5 Å 1ukrA (*Aspergillus niger*, meso) 2.4 Å 1xnb (*Bacillus circulans*, meso) 1.49 Å 1xnd (*Trichoderma harzianum*, meso) 1.8 Å 1xyn (*Trichoderma reesei*, meso) 2.0 Å 1yna (*Thermomyces lanuginosus*, 45°C) 1.55 Å
16. Xylanase (II) (Xyl-2)	1clxA (*Pseudomonas fluorescens*, meso) 1.8 Å 2exo (*Cellulomonas fimi*, meso) 1.8 Å 1xyzA (*Clostridium thermocellum*, 60°C) 1.4 Å
17. TATA box binding protein (TATA-BP)	1cdwA (Human, meso) 1.9 Å 1vokA (*Arabidopsis thaliana*, meso) 2.1 Å 1pczA (*Pyrococcus woesei*, 100°C) 2.2 Å
18. Adenylate kinase (ADK)	1ak2 (Bovine, meso) 1.92 Å 2ak3A (Bovine, meso) 1.85 Å 1aky (Yeast, meso) 1.63 Å 1ukz (Yeast, meso) 1.9 Å 1akeA (*Escherichia coli*, meso) 1.9 Å 3ukd (*Dictyostelium discoideum*, meso) 1.9 Å 1zip (*Bacillus stearothermophilus*, 52.5°C) 1.85 Å
19. Carboxypeptidase (CP)	2ctc (Bovine, meso) 1.4 Å 1nsa (Pig, meso) 2.3 Å 1pca (Pig, meso) 2.0 Å 1obr (*Thermoactinomyces vulgaris*, 55°C) 2.3 Å
20. Ornithine carbamoyltransferase (OCT)	2otcA (*Escherichia coli*, meso) 2.8 Å 1a1s (*Pyrococcus furiosus*, 100°C) 2.7 Å
21. Pyrophosphatase (PPase)	1obwA (*Escherichia coli*, meso) 2.15 Å 2prd (*Thermus thermophilus*, 72.5°C) 2.0 Å
22. CheY protein (CheY)	3chy (*Escherichia coli*, meso) 1.66 Å 2chf (*Salmonella typhimurium*, meso) 1.8 Å 1tmy (*Thermotoga maritima*, 80°C) 1.9 Å
23. Glutathione / trypanothione reductase (G/T reduct.)	1aogA (*Trypanosoma cruzi*, meso) 2.3 Å 1febA (*Crithidia fasciculata*, meso) 2.0 Å 1gerA (*Escherichia coli*, meso) 1.86 Å 3grs (Human, meso) 1.54 Å 1ebdA (*Bacillus stearothermophilus*, 52.5°C) 2.6 Å
24. Phosphofructokinase (PFK)	1pfkA (*Escherichia coli*, meso) 2.4 Å 4pfk (*Bacillus stearothermophilus*, 52.5°C) 2.4 Å
25. Triacylglycerol acylhydrolase (TAGAH)	1lgyA (*Rhizopus niveus*, meso) 2.2 Å 1tib (*Humicola lanuginosa*, 50°C) 1.84 Å 3tgl (*Rhizomucor miehei*, 45°C) 1.9 Å