Contribution of electrostatic interactions, compactness and quaternary structure to protein thermostability: lessons from structural genomics of *Thermotoga maritima*

Marc Robinson-Rechavi [1,3,4], Andreu Alibés [2,5], Adam Godzik [1,2]

1: Joint Center for Structural Genomics
University of California, San Diego
9500 Gilman Drive
La Jolla, CA 92093-0527
USA

2: Burnham Institute
10901 North Torrey Pines Road
La Jolla, CA 92037
USA

3: To whom correspondence should be addressed: e-mail, marc.robinson-rechavi@unil.ch; fax, +41 21 692 4165.

4: Present address: Department of Ecology and Evolution, Biophore, University of Lausanne, 1015 Lausanne, Switzerland.

5: Present address: Bioinformatics Unit, Centro Nacional de Investigaciones Oncológicas (CNIO) Melchor Fernández Almagro 3, 28029 Madrid, Spain

Running head: Structural determinants of protein thermostability

Summary

Studies of the structural basis of protein thermostability have produced a confusing picture. Small sets of proteins have been analyzed from a variety of thermophilic species, suggesting different structural features as responsible for protein thermostability. Taking advantage of the recent advances in structural genomics, we have compiled a relatively large protein structure dataset, that was also very carefully and selectively constructed; that is, the dataset contains only experimentally determined structures of proteins from one specific organism, the hyperthermophilic bacterium *Thermotoga maritima,* and those of close homologs from mesophilic bacteria. In contrast to the conclusions of previous studies, our analyses show that oligomerization order, hydrogen bonds, and secondary structure play minor roles in adaptation to hyperthermophily in bacteria. On the other hand, the data exhibit very significant increases in the density of salt bridges and in compactness for proteins from *T. maritima*. The latter effect can be measured by contact order or solvent accessibility, and in addition network analysis shows a specific increase in highly connected residues in this thermophile. These features account for changes in 96% of the protein pairs studied. Our results provide a clear picture of protein thermostability in one species, and a framework for future studies of thermal adaptation.

Keywords: thermostability, salt bridges, accessible surface area, evolution, network analysis

Introduction

What makes proteins from hyperthermophilic organisms stable and functional at temperatures higher than 80ºC [1], whereas most proteins are denatured at much lower temperatures? Studies of proteins from thermophiles and hyperthermophiles have not provided a clear answer, but rather an array of hypotheses [2; 3]. Yet, understanding the molecular basis of stability in thermophiles is necessary both to our fundamental understanding of protein structure and to prospects of engineering more thermostable proteins [4]. Applications of such engineering concern such diverse industries as household cleaning, textile washing, paper bleaching, fruit juice clarification or the digestibility of chicken food [5].

The most common approach for investigating the structural basis of protein thermostability has been the in-depth study of specific protein families. This type of analysis provides essential insight into specific mechanisms, but it is always difficult to judge how general the conclusions are. For example, different studies have highlighted the importance of salt bridges, solvent accessibility, hydrogen bonds, or dimerization surfaces, respectively. Availability of multiple genomes of thermophilic organisms allows a large-scale analysis of protein sequences, with the statistically significant identification of the most important features. Thus, the analysis of all predicted proteins from 71 genomes has shown that the most discriminating factor between the proteins from the mesophilic and thermophilic groups is the difference in charged–polar amino-acids (the "*CvP* bias") [6; 7]. Unfortunately, the potential impact of such analyses on our understanding of the structural aspects of thermostability is limited by our poor knowledge of the relationship between protein sequence and structure. Several studies tried to analyze the structures of thermostable proteins, but these studies were limited due to the use of either a small sample size [8; 9], or of comparative predicted and experimental structures to increase the sample size [10; 11; 12]. One large study compared 127 homologous PDB structures of thermophiles (defined as having a growth temperature above 65ºC) and mesophiles, but the study focused only on charged surface residues [13].

It is common in such statistical studies that proteins from different thermophilic and mesophilic species, and even different domains of life, be bundled in the analysis.

Since most thermophiles are archea, and most mesophiles are bacteria and eukaryotes, the comparisons are typically made between mostly archeal thermophiles, on the one hand, and a mix of eukaryotic and bacterial mesophiles, on the other. As a result, it is difficult to distinguish between features related to thermostability and those related to specific kingdoms of life.

In this study, we take advantage of the recent increase in structural data from *Thermotoga maritima*, a hyperthermophilic bacterium with an optimum growth temperature of 80ºC. A structural genomics effort on *T. maritima* [14], and related work in other Protein Structure Initiative centers, have contributed many of these structures, and made *T. maritima* one of the organisms with the best structural coverage of its genome. After removing structures from paralogous *T. maritima* genes, as well as those without a close homolog in a mesophilic bacteria, we have a dataset of 94 pairs of protein structures. This dataset has enabled us to characterize in detail the features that distinguish *T. maritima* proteins from their mesophilic homologs.

Results

Homology relations and structural comparison

The 94 pairs of structures in our dataset sample 9% of the gene families (excluding orphans) defined for the *Thermotoga maritima* genome in Hogenom [15], a database of gene families from complete genome sequences. Half of the protein structures from *T. maritima* analyzed were obtained in a targeted structural genomic effort [14].

We separated protein pairs according to their evolutionary relationship, first focusing on orthologs, which diverge by speciation, and usually perform the same function in both species. Thus, differences between them should be due mostly to differences between the species compared, i.e. here the *T. maritima* adaptation to thermophilic life style. Paralogs, on the other hand, diverge by gene duplication, and often differ in function, which in turn may impact the evolution of their structures. Despite this difference in relationship, for all structural features studied here, the variation is consistent between the two types of homologous pairs (Table 1). In particular, for all features with significant variation among homologs, the variation among paralogs is consistent with that variation found among orthologs, but exhibits greater average amplitude. This observation indicates that paralogs have diverged more than orthologs, but the increased divergence did not disturb the overall trend of the structural adaptation to high temperature. The increase in divergence between paralogs can be seen in the level of structural difference between homologous structures: the average root mean square distance (rmsd) is 2.05 Å between orthologs and 2.13 Å between paralogs. This interesting result supports the often made, but to our knowledge never verified in a large-scale analysis, statement about larger structural divergence of paralogs vs. orthologs.

On the other hand, very divergent paralogous pairs, detected not by BLAST but only by fold recognition (FFAS) [16], do not follow these statistical trends (data not shown). In this case, it is most probable that major changes in function are dominant over any trends associated with thermostability.

Pairwise comparison of *T. maritima* proteins vs. their mesophilic homologs

Many structural features were discussed in literature as possibly connected with protein thermostability. The list of features we tested is presented, with exact definitions and procedures for their calculations, in the Material and Methods section. Numerical values of all the structural features considered in this work were calculated for all protein chains in the dataset and compared by a pairwise t-test over homologous pairs. The results are presented in Table 1 and will be discussed in the following sections of the manuscript. All the features were chosen because of reasonable expectations that they would change in relation to thermostability, based on previously published analyses. Since in all cases the variation in paralogous pairs is consistent with that in orthologous pairs, the most representative results appear to be those of the combined dataset, which represents the largest quantity of data (last column of Table 1).

Although there have been reports of higher-order quaternary structure in thermophiles [17], this is the one feature which shows no directionality between *T. maritima* and mesophiles. All other features show some measure of change in the direction expected if thermostable proteins are more compact and have more bonds. Because of test repetition, however, not all of these changes can be considered statistically significant. At the other extreme, there are five features for which differences between *T. maritima* and mesophiles are highly significant (results in bold in Table 1).

Consistent with the expectation that *T. maritima* proteins are more stable [18; 19; 20], one of the most significant differences with proteins from mesophiles is lower empirical estimates of stabilization energy [21; 22]. Although all components of the energy (burial, local and contact energy) are lower in *T. maritima* on average (not shown), the difference is especially strong for contact potential (mean difference = -0.022, $p < 10^{-4}$). Empirical energy functions, similar to those used here, have been used to estimate quality of protein models [23; 24], stability of point mutations [25] and to design protein sequences for artificial proteins [26], but only recently were used to design thermostable mutations [27]. An interesting observation is that, while there is significantly lower empirical stabilization energy in *T. maritima* for pairs of homologous proteins, there is also a large overlap between the *T. maritima* and mesophilic distributions. Thus a quarter of our protein sample from mesophilic bacteria have lower empirical stabilization energy than the median for the proteins from *T. maritima*, although the latter are presumably all more thermostable. This may indicate that the empirical estimate of energy difference between homologs (i.e. very similar structures) is more accurate than the estimate of absolute value of the stabilization energy, a common situation for estimation of physical quantities.

Another highly significant difference between *T. maritima* and mesophiles is the number of salt bridges relative to protein length, as often suggested in literature [2; 13; 18; 19; 28; 29]. Consistent with the hypothesis of Suhre and Claverie [7] concerning the role of charged residues, a change in *CvP* bias (0.063 in mesophiles vs. 0.14 in *T. maritima*, $p < 10^{-4}$) is positively correlated to the enrichment in salt bridges (Figure 1A), although this explains only 7% of the variance. While cation-π interactions were previously suggested to play an important role in thermostability [11], the difference between *T. maritima* and mesophiles appears only marginally significant (Table 1). There is also a marginally significant correlation between enrichment in salt bridges and in cation-π interactions in *T. maritima* proteins (Figure 1B). Both salt bridges and cation-π interactions are electrostatic interactions which stabilize protein structure, especially at high temperature [30]. Taken together, these observations suggest that cation-π interactions may play a role in stabilizing *T. maritima* proteins, although they seem to be of lesser importance than salt bridges. Together, both types of electrostatic interactions form an average of 13.1 bonds per *T. maritima* protein chain, compared to 10.8 per chain in homologs from mesophiles.

Compactness, an important but complex feature

The other features which differ very significantly between *T. maritima* and mesophiles are all related to protein compactness. From considerations on an earlier (and smaller) dataset, we reported recently one measure of more compact proteins in *T. maritima* [31], higher contact order, which has been shown to be related to folding rate [32]. This effect is even stronger on a larger dataset, and in addition we observe that thermophilic proteins also have significantly lower relative accessible surface area, thus are less accessible to the solvent, which happens to be very hot water.

As structure compactness is a complex notion, we have endeavored to characterize this difference further. For this, we have used a newly developed method of protein structure analysis, where a structure is described as a network of interacting nodes (see Methods) [33; 34] (Alibés and Godzik in preparation). Network analysis is an established framework suitable for dissecting large sets of interacting data to obtain both global and local properties. A tool with long traditions in engineering, it recently become popular in the analysis of biologically related data [35], even though most analyses have concentrated on networks of protein interactions [36], metabolic networks [37], or networks of domain connections [38]. The main network properties usually analyzed are the average connectivity, k, which is the average number of links per node; the mean geodesic (shortest) distance between node pairs, L; and the clustering coefficient, C, which is the average ratio between the number of links among neighbors of a node and the maximum possible number of such links. Intuitively, more compact networks should be characterized by higher connectivity, lower mean distance between nodes, and higher clustering coefficient. On a global level, all network properties investigated are indeed consistent with *T. maritima* proteins being more compact, with significant differences in connectivity (Table 1) and in clustering coefficient C (0.299 vs. 0.288, p = .0005). For example, despite extremely similar structures, the more compact HAM1 homolog from *T.*

maritima has visibly more connections in its network representation than the ortholog from *E. coli* (Figure 2). We have also evaluated the number of all 3- and 4-node subgraphs. The number of such subgraphs per node is higher in *T. maritima* than in mesophilic homologs for each subgraph (not shown); this finding points again to higher compactness of *T. maritima* proteins.

The difference of average connectivity in the network between homologs could be achieved by at least three different mechanisms: an evenly distributed increase in connectivity over all residues; a decrease in the number of residues with few connections; or an increase in the number of highly connected residues. The comparison of connectivity distribution in *T. maritima* proteins and their mesophilic homologs (Figure 3) shows that it is the latter mechanism that is responsible for the difference we observe: there is a clear increase in highly connected residues (7-10 neighbors per node).

There are other characteristics of *T. maritima* proteins whose variation is of marginal significance, but are also indicative of the manner in which compactness is achieved (Table 1). *T. maritima* proteins are thus shorter on average than proteins from mesophiles, and have a lower proportion of sites in disorganized regions of the structure (i.e. neither helices nor sheets). Variation in these two features is surprisingly not correlated: the length difference is not simply due to the loss of disorganized regions. They are each correlated to variation in other features, such as accessible surface area or contact order (not shown). The slight increase in hydrogen bond density also observed appears to be mostly due to the loss of disorganized regions ($r = -0.72$, $p < 10^{-4}$), and thus it appears improbable that an enrichment in hydrogen bonds play an important role in thermostability *per se*, as previously suggested [39].

Relation between compactness, electrostatic interactions, and quaternary structure

Overall, there are five features of proteins whose variation can be related to an increase in protein compactness, and which are correlated to each other to some degree (i.e. Figure 1C): contact order, accessible surface area, connectivity, protein length, and proportion of disorganized regions. There also appears to be a relation between the variation in some of these features and the evolution of quaternary structure. Proteins with the same quaternary structure in *T. maritima* and the mesophile on average do not change accessible surface area (diff = -0.014, *p* = .41), while the average decrease seen on the whole set is due to the proteins whose quaternary structure does vary between species (diff = –0.077, *p* = .0002). While the decrease in accessible surface area is correlated to the increase in connectivity over all proteins (Figure 1C), it is correlated to the increase in contact order only for these proteins of variable quaternary structure (Figure 1D). Similarly, the increase in contact order and in connectivity are correlated for proteins of variable quaternary structure (r = 0.36, *p* = .028), but not for those of constant quaternary structure (r = 0.016, *p* = .92). These observations do not depend on how the quaternary structure changes: proteins that form larger or smaller complexes in *T. maritima* behave in the same manner. Thus despite the lack of directional change in quaternary structure between *T. maritima* and mesophiles, such change does play an important role in the evolution of thermostability. It should be noted that the increase in either connectivity or contact order is independent of quaternary structure variation.

There is no correlation at all between variation in any of the features describing compactness, and any of the features describing electrostatic interaction. Yet a behavior mirroring somewhat that of accessible surface area is observed for salt bridge variation: the increase is strongest for proteins with conserved quaternary structure (diff = 0.013, *p* = .0004), for which there is also a good correlation with variation in cation-π interactions (r = 0.31, *p* = .040). Whereas for proteins with variable quaternary structure the increase in salt bridge density is less important (diff = 0.0077, *p* = .015), and not correlated with variation in cation-π interactions (r = 0.17, *p* = .28).

A consequence of the independent variation of features related to compactness or electrostatic interactions is that only three *T. maritima* proteins out of 94 are neither more compact by some measure, nor have more electrostatic interactions. Interestingly, all features compared vary independently of the variation in empirical stabilization energy, and the three proteins without any apparent increase in compactness nor electrostatic interactions all have lower empirical energy in *T. maritima*. These proteins are a 33 kDa chaperonin (PDB: 1vq0), a transcriptional regulator of the CrP family (PDB: 1o51), and a zinc-containing alcohol dehydrogenase (PDB: 1vj0). There is no obvious connection between these proteins, which may be stabilized by mechanisms which still escape us.

T. maritima **is representative of thermophiles**

Although proteins from thermophilic organisms are often compared to those from mesophilic organisms under the *a priori* assumption that observed differences are due to thermostability [40], this assumption is far from obvious. To verify it, we compared *T. maritima* proteins to homologs from other thermophiles (Table 2). There appears to be little difference. Thus the differences that we observe with mesophiles (Table 1) are very likely due to adaptation to high temperature in *T. maritima* and not to any specific *T. maritima* features. If anything, homologs from other thermophiles carry the same trends as *T. maritima* further. This might arise because several have even higher growth temperatures, but data is insufficient to test correlations with exact growth temperatures.

There are 22 *T. maritima* proteins for which we have homologous structures both from a mesophile and another thermophile. For all of the features considered is there is no correlation between the other-thermophile - mesophile difference, and the *T. maritima* - mesophile difference (not shown), which suggests that different thermostability strategies are used in different species for homologous proteins. To verify this, we looked into eight protein families for which structures are available from at least two other thermophiles, in addition to a mesophilic bacteria and *T. maritima*. In only one case, ornithine carbamoyltransferase (*T. maritima* PDB: 1vlv), are all trends consistent in the two other thermophiles sampled, although both are archeal and one is a paralog (PDB: 1a1s and 1ml4). In all other families, homologs from different species follow different strategies. For example, the *T. maritima* cell division protein FtsY (PDB: 1vma) is more compact than its ortholog from *E. coli* (PDB: 1fts) (accessible surface area difference = -0.79; contact order difference = 0.010), but has only slightly more salt bridges (difference = 0.0036). Yet homologs from four other species, including an ortholog from the bacteria *Thermus aquaticus* (PDB: 1okk), are all strongly enriched in salt bridges (difference = 0.0069 to 0.036), but less compact than the *E. coli* protein. Different features can thus contribute to thermostability not only in different proteins, but also in different species.

Thus for each protein, mechanisms of thermostability may be different between species, or between paralogs, but the average behavior of proteins is similar in different thermophiles, of which *T. maritima* is representative: more electrostatic interactions, more compact.

Discussion

Although studies of individual proteins have yielded confusing results concerning the general causes of thermostability, we have been able to define structural features which distinguish proteins from the hyperthermophile *T. maritima* from their homologs in mesophiles with strong statistical significance. Although these features result from statistical computations, not direct experimentation, they are based on high quality structures, and we believe the conclusions to be relatively robust. These results were made possible in large part by the recent progress of structural genomics (e.g. [14]).

The features which distinguish proteins between *T. maritima* and mesophiles concern two broad types of properties: *T. maritima* proteins are on average more compact, and they have on average more electrostatic interactions. The high number of salt bridges in proteins from thermophiles, which is the most significant feature in our comparison, may also be the one feature of thermostable proteins which has been consistently noticed in various studies [2; 8; 18; 19; 41]. Thus, in a comparison of 13 genomes, more salt bridges were found in predicted protein sequences from thermophiles, especially inside predicted helices [10]. A similar result was found using PDB structures to model 125 large families of homologous proteins from 30 genomes [11]. More electrostatic interactions in proteins from thermophiles were also found in several comparisons of homologous experimental structures [8; 9; 13; 28]. While most studies focus on salt bridges, Chakravarty and Varadarajan [11] also found more cation–π interactions. Our results suggest that they may play a role, but much smaller than salt bridges. Both cation–π and salt bridges are electrostatic interactions that are stabilizing at high temperatures, even in those proteins for which such bridges may be destabilizing at room temperature due to changes in dielectric response [42]. Halogen–π interactions may also be important in proteins [43], but the extant data are too limited to test the significance of this effect in the stability of thermophilic proteins. A very recent study has highlighted the role of disulfide bonds [12]. The methodology used here did not allow us to confirm or infirm this observation (not shown), but we note that *T. maritima* is predicted to have fewer disulfide bonds than most other hyperthermophiles in Figure 2 of Beeby et al [12].

In parallel with the strong difference in electrostatic interactions for structures, the most discriminating factor between sequences from mesophiles and thermophiles is a bias in polar amino acids of the proteins, the "*Charged vs. Polar (CvP)* bias" [6; 7]. The correlation we find between the increase in electrostatic interactions and in *CvP* bias is consistent with the suggested structural role of this difference in composition [7]. Yet the correlation only explains 7% of the variance. A simple explanation is that the position of amino acid changes in the structure is more important than a global increase in charged residues. In a similar manner, the loss of sites in disorganized regions of the structure accounts for less than 5% of the variance in accessible surface area; such loss does not seem as important as previously thought [44]. We also reported previously the absence of correlation between disorganized regions and increased contact order [31]. Network analysis on the other hand allows us to pin down which sites contribute most significantly to the increased compactness of *T. maritima* proteins (Figure 3): those which are already strongly connected. So compactness may come less from "tightening the loops", which would show in a large contribution from the loss of disorganized regions, and increased connectivity of the less connected residues, but more from an even better connectivity in those protein regions which already have a tendency to compactness. This suggests that it may be less disruptive to increase compactness in regions whose functional role is already consistent with high compactness, while conserving the properties of low connectivity regions, which may play functional roles such as protein flexibility.

We have reported recently that increased compactness was detected by higher contact order in proteins from *T. maritima*, compared to mesophiles [31]. We also observe a very significant difference in relative accessible surface area, which is consistent with some previous studies [11], but in contradiction to others [9]. Of note, Berezovsky et al. [29] found more compact proteins in *Pyrococcus* but not *T. maritima*, in contrast with our observations. An interesting observation is that the way proteins become more compact differs according to the evolution of quaternary structure. For the half of the proteins studied which have the same quaternary structure in the species compared, solvent accessibility does not change significantly, while higher compactness is achieved all the same, as shown by higher contact order and connectivity. These proteins also tend to have the strongest increase in electrostatic interactions. In the other half of the proteins, changes in quaternary structure are coupled with a decrease in solvent accessibility, which is correlated to other measures of compactness, notably higher contact order. This can be achieved both through lower-order or higher-order complexes; in contradiction with some previous reports [17], we did not find any specific trend for higher-order oligomers in *T. maritima*. It is possible, of course, that such a trend exists in some other thermophilic species, although it was also not found in a study including a mix of various species of bacteria and of archea [9]. We suggest that modifications in quaternary structure are favored when they result in a decrease in relative solvent accessibility. It is possible that the change in quaternary structure allows an increase in hydrophobic interactions despite their lower efficiency at high temperature.

A recent success of protein engineering has been the thermostabilization of yeast cytosine deaminase [27]. This is a specifically interesting example because experimental structures of the wild type and the mutants are available, and because enzymatic function was maintained. Consistent with the observation of 70 Å^2 more buried surface area [27], solvent accessibility is lower in the engineered protein. But we also observe a regular increase of contact order, of connectivity, and of clustering coefficient, from the wild type to the most stable triple mutant, with intermediate values in the double mutant of intermediate thermostability. Thus, protein compactness appears quite relevant to the engineering of more thermostable proteins, as also shown by the artificial peptide BBAT1 [31; 45].

Different strategies to achieve thermostability could be expected among different proteins, since they have different constraints related to structure and function. Indeed, this observation has already been made multiple times [2; 8; 12; 19; 29; 40; 41; 46; 47]. Our results confirm this diversity, with two major properties of protein structure that vary independently from each other. Different strategies to achieve thermostability are also found between different organisms for the same protein family. Variation of each feature in other thermophilic species is not correlated to the variation in *T. maritima*. And for the few proteins with structures solved in at least three thermophiles, including *T. maritima*, and a mesophile, different variation of the features studied is found in different species.

This may pose a problem for studies that average over different thermophilic species in each protein family [11], since different strategies may compensate each other in the calculation.

Structural genomics of *T. maritima* provides a powerful tool to investigate the structural basis of protein thermostability. We have identified factors that are directly related to thermophily, since they do not vary among thermophiles but do vary between thermophilic and mesophilic bacteria. Moreover, the same type of variation is observed among orthologs and close paralogs. We believe this study provides a clear test of different structural features that have been proposed to correlate with thermophily.

Materials and Methods

Sequence analysis

All sequences corresponding to protein structures were recovered from the PDB [48] on 14 June 2005. The subset of entries from *Thermotoga maritima* was compared by BlastP to all other sequences. For each *T. maritima* entry that had at least one hit with E-value under e-4, aligned homologous proteins from completely sequenced genomes were recovered from Hogenom at PBIL [15; 49]. These alignments were edited to add sequences of homologous PDB entries that are absent from Hogenom (typically from organisms whose genome is not sequenced) and to merge protein families that were classified separately in Hogenom, but were homologous according to the results of Blast on the PDB.

For each alignment, a phylogenetic tree was built by PhyML [50], with the JTT model and a gamma distribution between sites (parameter alpha estimated by PhyML with 8 categories). These trees were used to assess homology relations between PDB entries: orthology, paralogy, or xenology (horizontal transfer). Five cases of suspected horizontal transfer were excluded from the final dataset. For 10 cases where more than one *T. maritima* paralog from the same family was available, only one was used.

The dataset used includes 94 chains from *T. maritima* proteins of known tertiary structure, of which 62 have a mesophilic bacterial ortholog of known structure and 32 have a mesophilic bacterial close paralog of known structure.

Classical structure analysis

Structures were recovered from the PDB [48], and single chains were extracted. For homo-oligomers, only chain A was used in calculations. For hetero-oligomers, chains were treated separately according to their homology relations as established by phylogenetic analysis. Structures that cover different portions of homologous proteins were not used for comparisons. For example, the available structure for the *T. maritima* chemotaxis protein (1b3q) only covers the C-terminal of the protein, while the available structure from the *Salmonella typhimurium* chemotaxis protein (1i5n) only covers the N-terminal of the protein. When there were several mesophilic bacterial structures orthologous to a same *T. maritima* structure, the one with the lowest RMSD to *T. maritima* was used. The same was done for paralogous structures.

For each chain analyzed, the following features were computed, excluding all HET atoms (e.g., water, cofactors):

- the length of the protein sequence reported by Swissprot-TrEMBL [51].
- the ratio accessible surface area/sphere surface, with the sphere surface calculated as the surface of a sphere of the same volume as the protein, as in Kumar et al. [9]. Accessible surface area calculated by the program calc-surface [52], with a probe size of 1.4 Å, and the volume calculated by Voronoi approximation by the program calc-volume [53].
- salt bridges, calculated by the program WHATIF in its WWW implementation [54]; only bridges involving no HIS atom, between atoms less than 4 Å apart [55], were counted. For some structures, “bridges” with a distance of zero Å are reported; they were considered to be artifacts, and were not counted.
- energetically significant cation–π interactions, calculated by the program CaPTURE [56].
- hydrogen bonds, calculated by the program DSSP [57], adding all types of hydrogen bonds reported.
- proportions of residues in alpha helices, beta strands, or disorganized regions. Secondary structure attribution of residues was calculated by the program DSSP [57], following the classification of Chakravarty et al. [11].

- quaternary structure, predicted by PQS [58]; there is no prediction for structures solved by NMR.
- potential energy [21], calculated by the program PSQS (available at http://www1.jcsg.org/psqs/psqs.cgi).
- relative contact order, calculated by the program contactOrder.pl (http:// http://depts.washington.edu/bakerpg/contact_order/), which implements the definition of Plaxco et al. [32]: any non water atoms separated by less than 6 Å are considered "in contact".

Features were compared between homologous chains by a paired t-test. Divergence between homologous chains was measured by RMSD, calculated by FATCAT [59] without flexibility.

Network structure analysis

The 3D protein structures were translated into a network structure according to the following rules: (a) each residue corresponds to a node; (b) two nodes are linked if the two residues have any two atoms closer than 4.5 Å. These links are non-directed. Amino acids closer than 3 positions in the protein sequence are not considered linked in order to avoid trivial information. General properties of networks can then be calculated and the difference between each pair of proteins analyzed. This approach is similar to Greene and Highman [33], but in their analysis a cut-off value of 5 Å is used, links are weighted by the number of close atoms between each pair of amino acids, and use two different interactions: short-range (<10 positions apart in sequence) and long-range (≥10 positions). Weighting the interactions makes networks more dependent on the value of the cutoff. Our approach, which depends on less user-defined parameters, may provide a closer look at the differences between thermophilic and mesophilic proteins. The cut-off value of 4.5 Å is small enough to ensure capturing the small differences between thermophilic and mesophilic protein structures and large enough to obtain significant results with a sufficiently connected network. The number of all types of subgraphs of 3 or 4 nodes were computed using the program MFINDER [60], that has already been used to analyze various biologically relevant networks.

Acknowledgments

We thank Jaume Canaves, Iddo Friedberg, Marta Ibañes, Lukasz Jaroszewski, Dalit Rechavi, Robert Schwarzenbacher, John Wooley, and Yuzhen Ye for helpful suggestions. We thank all members of the Joint Center for Structural Genomics. This research was supported by NIH Protein Structure Initiative grant, number P50 GM62411.

References

1. Daniel, R. M. & Danson, M. J. (2001). Assaying activity and assessing thermostability of hyperthermophilic enzymes. *Methods Enz* 334, 283-293.
2. Petsko, G. A. (2001). Structural basis of thermostability in hyperthermophilic proteins, or "There's more than one way to skin a cat". *Methods Enz* 334, 469-478.
3. Sterner, R. & Liebl, W. (2001). Thermophilic adaptation of proteins. *Crit Rev Biochem Mol Biol* 36, 39-106.
4. Eijsink, V. G. H., Bjork, A., Gaseidnes, S., Sirevag, R., Synstad, B., Burg, B. v. d. & Vriend, G. (2004). Rational engineering of enzyme stability. *J. Biotechnol.* 113, 105-120.
5. Vieille, C. & Zeikus, G. J. (2001). Hyperthermophilic Enzymes: Sources, Uses, and Molecular Mechanisms for Thermostability. *Microbiol. Mol. Biol. Rev.* 65, 1-43.
6. Cambillau, C. & Claverie, J.-M. (2000). Structural and Genomic Correlates of Hyperthermostability. *J. Biol. Chem.* 275, 32383-32386.
7. Suhre, K. & Claverie, J.-M. (2003). Genomic Correlates of Hyperthermostability, an Update. *J. Biol. Chem.* 278, 17198-17202.
8. Szilagyi, A. & Zavodszky, P. (2000). Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits: results of a comprehensive survey. *Structure Fold Des* 8, 493-504.
9. Kumar, S., Tsai, C. J. & Nussinov, R. (2000). Factors enhancing protein thermostability. *Protein Eng* 13, 179-91.
10. Das, R. & Gerstein, M. (2000). The stability of thermophilic proteins: a study based on comprehensive genome comparison. *Funct Integr Genomics* 1, 76-88.
11. Chakravarty, S. & Varadarajan, R. (2002). Elucidation of factors responsible for enhanced thermal stability of proteins: a structural genomics based study. *Biochemistry* 41, 8152-61.
12. Beeby, M., O'Connor, B. D., Ryttersgaard, C., Boutz, D. R., Perry, L. J. & Yeates, T. O. (2005). The Genomics of Disulfide Bonding and Protein Stabilization in Thermophiles. *PLoS Biology* 3, e309.
13. Alsop, E., Silver, M. & Livesay, D. R. (2003). Optimized electrostatic surfaces parallel increased thermostability: a structural bioinformatic analysis. *Protein Eng.* 16, 871-874.

14. Lesley, S. A., Kuhn, P., Godzik, A., Deacon, A. M., Mathews, I., Kreusch, A., Spraggon, G., Klock, H. E., McMullan, D., Shin, T., Vincent, J., Robb, A., Brinen, L. S., Miller, M. D., McPhillips, T. M., Miller, M. A., Scheibe, D., Canaves, J. M., Guda, C., Jaroszewski, L., Selby, T. L., Elsliger, M. A., Wooley, J., Taylor, S. S., Hodgson, K. O., Wilson, I. A., Schultz, P. G. & Stevens, R. C. (2002). Structural genomics of the Thermotoga maritima proteome implemented in a high-throughput structure determination pipeline. *Proc Natl Acad Sci U S A* 99, 11664-9.

15. Kersey, P., Bower, L., Morris, L., Horne, A., Petryszak, R., Kanz, C., Kanapin, A., Das, U., Michoud, K., Phan, I., Gattiker, A., Kulikova, T., Faruque, N., Duggan, K., McLaren, P., Reimholz, B., Duret, L., Penel, S., Reuter, I. & Apweiler, R. (2005). Integr8 and Genome Reviews: integrated views of complete genomes and proteomes. *Nucl. Acids Res.* 33, D297-302.

16. Rychlewski, L., Jaroszewski, L., Li, W. & Godzik, A. (2000). Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Sci* 9, 232-41.

17. Backmann, J. & Schafer, G. (2001). Thermodynamic analysis of hyperthermostable oligomeric proteins. *Methods Enz* 334, 328-342.

18. 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, 1516-27.

19. Xiao, L. & Honig, B. (1999). Electrostatic contributions to the stability of hyperthermophilic proteins. *J Mol Biol* 289, 1435-44.

20. Kumar, S., Tsai, C. J. & Nussinov, R. (2001). Thermodynamic differences among homologous thermophilic and mesophilic proteins. *Biochemistry* 40, 14152-65.

21. Godzik, A., Kolinski, A. & Skolnick, J. (1995). Are proteins ideal mixtures of amino acids? Analysis of energy parameter sets. *Protein Sci* 4, 2107-17.

22. Pawlowski, K., Jaroszewski, L., Bierzynski, A. & Godzik, A. (1997). Multiple model approach--dealing with alignment ambiguities in protein modeling. *Pac Symp Biocomput*, 328-39.

23. Petrey, D. & Honig, B. (2000). Free energy determinants of tertiary structure and the evaluation of protein models. *Protein Sci* 9, 2181-91.

24. Dominy, B. N. & Brooks, C. L. (2002). Identifying native-like protein structures using physics-based potentials. *J Comput Chem* 23, 147-60.

25. Bordner, A. J. & Abagyan, R. A. (2004). Large-scale prediction of protein geometry and stability changes for arbitrary single point mutations. *Proteins* 57, 400-13.

26. Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L. & Baker, D. (2003). Design of a novel globular protein fold with atomic-level accuracy. *Science* 302, 1364-8.

27. Korkegian, A., Black, M. E., Baker, D. & Stoddard, B. L. (2005). Computational Thermostabilization of an Enzyme. *Science* 308, 857-860.

28. Kumar, S. & Nussinov, R. (2002). Close-range electrostatic interactions in proteins. *Chembiochem* 3, 604-17.

29. Berezovsky, I. N. & Shakhnovich, E. I. (2005). Physics and evolution of thermophilic adaptation. *Proc Natl Acad Sci U S A* 102, 12742-7.

30. Elcock, A. H. (1998). The stability of salt bridges at high temperatures: implications for hyperthermophilic proteins. *J Mol Biol* 284, 489-502.

31. Robinson-Rechavi, M. & Godzik, A. (2005). Structural genomics of *Thermotoga maritima* proteins shows that contact order is a major determinant of protein thermostability. *Structure* 13, 857-860.

32. Plaxco, K. W., Simons, K. T. & Baker, D. (1998). Contact order, transition state placement and the refolding rates of single domain proteins1. *J Mol Biol* 277, 985-994.

33. Greene, L. H. & Higman, V. A. (2003). Uncovering network systems within protein structures. *J Mol Biol* 334, 781-91.

34. Dokholyan, N. V., Li, L., Ding, F. & Shakhnovich, E. I. (2002). Topological determinants of protein folding. *Proc Natl Acad Sci U S A* 99, 8637-41.

35. Barabasi, A. L. & Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nat Rev Genet* 5, 101-13.

36. Jeong, H., Mason, H. P., Barabasi, A. L. & Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature* 411, 41-42.

37. Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N. & Barabasi, A. L. (2002). Hierarchical organization of modularity in metabolic networks. *Science* 297.

38. Ye, Y. & Godzik, A. (2004). Comparative analysis of protein domain organization. *Genome Res* 14, 343-53.

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

40. Jaenicke, R. & Bohm, G. (2001). Thermostability of proteins from Thermotoga maritima. *Methods Enz* 334, 438-469.

41. Rees, D. C. (2001). Crystallographic analyses of hyperthermophilic proteins. *Methods Enz* 334, 423-437.

42. Dominy, B. N., Minoux, H. & Brooks, C. L., 3rd. (2004). An electrostatic basis for the stability of thermophilic proteins. *Proteins* 57, 128-41.

43. Saraogi, I., Vijay, V. G., Das, S., Sekar, K. & Guru Row, T. N. (2003). C-halogen...[pi] interactions in proteins: a database study. *Crystal Engineering* 6, 69-77.

44. Thompson, M. J. & Eisenberg, D. (1999). Transproteomic evidence of a loop-deletion mechanism for enhancing protein thermostability. *J Mol Biol* 290, 595-604.

45. Ali, M. H., Peisach, E., Allen, K. N. & Imperiali, B. (2004). X-ray structure analysis of a designed oligomeric miniprotein reveals a discrete quaternary architecture. *Proc Natl Acad Sci U S A* 101, 12183-12188.

46. Cowan, D. A. (1997). Thermophilic proteins: Stability and function in aqueous and organic solvents. *Comp Biochem Physiol A: Physiol* 118, 429-438.

47. Hollien, J. & Marqusee, S. (2002). Comparison of the folding processes of *T. thermophilus* and *E. coli* Ribonucleases H1. *J Mol Biol* 316, 327-340.

48. Bourne, P. E., Addess, K. J., Bluhm, W. F., Chen, L., Deshpande, N., Feng, Z., Fleri, W., Green, R., Merino-Ott, J. C., Townsend-Merino, W., Weissig, H., Westbrook, J. & Berman, H. M. (2004). The distribution and query systems of the RCSB Protein Data Bank. *Nucl. Acids. Res.* 32, D223-225.

49. Perrière, G., Combet, C., Penel, S., Blanchet, C., Thioulouse, J., Geourjon, C., Grassot, J., Charavay, C., Gouy, M., Duret, L. & Deléage, G. (2003). Integrated databanks access and sequence/structure analysis services at the PBIL. *Nucl. Acids. Res.* 31, 3393-3399.

50. Guindon, S. & Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 52, 696-704.

51. Bairoch, A. & Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucl. Acids. Res.* 28, 45-8.

52. Gerstein, M. (1992). A Resolution-Sensitive Procedure for Comparing Protein Surfaces and its Application to the Comparison of Antigen-Combining Sites. *Acta Crystallogr A* 48, 271-276.

53. Tsai, J., Taylor, R., Chothia, C. & Gerstein, M. (1999). The packing density in proteins: standard radii and volumes. *J Mol Biol* 290, 253-66.

54. Rodriguez, R., Chinea, G., Lopez, N., Pons, T. & Vriend, G. (1998). Homology modeling, model and software evaluation: three related resources. *Bioinformatics* 14, 523-8.

55. Kumar, S. & Nussinov, R. (2002). Relationship between ion pair geometries and electrostatic strengths in proteins. *Biophys J* 83, 1595-612.

56. Gallivan, J. P. & Dougherty, D. A. (1999). Cation-pi interactions in structural biology. *Proc Natl Acad Sci U S A* 96, 9459-9464.

57. Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22, 2577-637.

58. Henrick, K. & Thornton, J. M. (1998). PQS: a protein quaternary structure file server. *Trends Biochem Sci* 23, 358-61.

59. Ye, Y. & Godzik, A. (2003). Flexible structure alignment by chaining aligned fragment pairs allowing twists. *Bioinformatics* 19 Suppl 2, II246-II255.

60. Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D. & Alon, U. (2002). Network motifs: simple building blocks of complex networks. *Science* 298, 824-7.

61. DeLano, W. L. (2002). The PyMOL Molecular Graphics System. DeLano Scientific LLC, San Carlos, CA, USA.

Figure 1: Correlations between variations in structural features

All correlations among differences between values for proteins from *T. maritima* and their homologs from mesophiles.

(A) Correlation between difference in salt bridge density and in *CvP* bias.

(B) Correlation between difference in salt bridge density and in cation-π interaction density.

(C) Correlation between difference in connectivity (k) and in relative solvent accessible surface area (rel. ASA).

(D) Correlation between difference in relative contact order (CO) and in relative solvent accessible surface area (rel. ASA). In red, significant correlation for proteins with the same quaternary structure in both species (N = 43). In blue, non significant correlation for proteins with different quaternary structures in the two species (N = 43). For 8 pairs one or the other structure was solved by NMR, preventing the prediction of quaternary structure by PQS [58].

Figure 2: HAM1 protein homolog, an example of higher connectivity in *T. maritima*

In red, *T. maritima*; in blue, the ortholog from *E. coli*. Left, the structure of chain A of each structure (1vp2 is a tetramer, 1k7k is a dimer). Right, the same structures translated into networks; spheres correspond to residues from the chains, ordered according to the sequence, which are the nodes of the network; lines correspond to the links of the network. The structures were represented using PyMol [61]; networks were represented using Agna 2.1.1 (http://www.geocities.com/imbenta/agna/).

Figure 3: Distribution of connectivity values

Distributions of connectivity values for all residues of all proteins compared. Red curve: residues from *T. maritima* proteins; blue curve: residues from mesophilic proteins. Y axis: number of connections for each value of *k* (*k* connections times *N(k)* residues), normalized by the total number of residues (*N*).