ABSTRACT

We address the question of the thermal stability of proteins in thermophiles through comprehensive genome comparison, focussing on the occurrence of salt bridges. We compared a set of twelve genomes (from four thermophilic archaeons, one eukaryote, six mesophilic eubacteria, and one thermophilic eubacteria). Our results showed that thermophiles have a greater content of charged residues than mesophiles, both at the overall genomic level and in alpha helices. Furthermore, we found that in thermophiles the charged residues in helices tend to be preferentially arranged with a 1-4 helical spacing and oriented so that intra-helical charge pairs agree with the helix dipole. Collectively, these results imply that intra-helical salt bridges are more prevalent in thermophiles than mesophiles and thus suggest that they are an important factor stabilizing thermophilic proteins. We also found that the proteins in thermophiles appear to be somewhat shorter than those in mesophiles. However, this later observation may have more to do with evolutionary relationships than with physically stabilizing factors. In all our statistics we were careful to take into account and control for various biases. These could have, for instance, arisen due to repetitive or duplicated sequences. In particular, we repeated our calculation using a variety of random and directed sampling schemes. One of these involved making a "stratified sample," a representative cross-section of the genomes derived from a set of 52 orthologous proteins present roughly once in each genome. For another sample, we focused on the subset of the 52 orthologs that had a known 3D structure. This allowed us to determine the frequency of tertiary as well as mainchain salt bridges. Our statistical controls supported our overall conclusion about the prevalence of salt bridges in thermophiles in comparison to mesophiles. Supplementary information is available from http://bioinfo.mbb.yale.edu/genome/thermophile.

INTRODUCTION

What are Thermophiles and how do their Proteins Achieve Stability?

The archaea and a few eubacteria, commonly known as thermophiles, thrive in high temperatures. They live in places such as hot springs and deep-sea hydrothermal vents under extreme conditions. It is not well understood how thermophiles stabilize proteins at the elevated temperatures that denature normal-temperature (10 to 45°C) mesophilic proteins. So far, several biophysical studies have been performed to determine the stability factors. These studies suggest about fifteen different physicochemical factors for thermostability, such as hydrogen bonding, hydrophobic internal packing, helix dipole stabilization, and salt bridge optimization among others. The factors have been reviewed by several authors (Gupta, 1995; Querol *et al*., 1996; Russel *et al*., 1995; Vieille *et al*., 1996; Colacino & Crichton, 1997; Jaenicke & Bohm, 1998; Scandurra *et al.*, 1998; Vogt *et al*., 1997a,b).

The Salt Bridge as a Major Factor

Since the study of thermostability of bacterial ferredoxins by Perutz (Perutz & Raidt, 1975), a large number of 3D structures of thermophilic proteins have been determined. These structures as well as structural information obtained through homology modelling revealed that there is a strong correlation between the number of salt bridges and protein thermal stability (Wallon *et al*., 1997; Hennig *et al*., 1995, 1997; Knapp *et al*., 1997; Korndorfer *et al*., 1995; Russell *et al*., 1997; Salminen *et al*., 1996; Yip *et al*., 1995; Auerbach *et al*., 1998; Xiao & Honig, 1999; Spassov *et al.*, 1995; Szilagyi & Zavodszky, 1995; Wallon *et al.*, 1997). A theoretical study by Elcock (1997) further supported this correlation, and showed that since the hydration free energies of the charged groups become less favorable at high temperatures, the unfavourable desolvation penalty incurred on forming the salt bridges is reduced in magnitude. Therefore, using the logic of solvation, the study suggested that the salt bridge becomes more stable at elevated temperatures.

Studies of protein structures have shown that there are several ways in which salt bridges can stabilize proteins. Ion pair networks, helix stabilizing salt bridges, salt bridges buried in a hydrophobic core and surface salt bridges between two subunits are among the most frequently encountered types (Petukhov *et al*., 1997; Hennig *et al*., 1995; Korndorfer *et al*., 1995; Lebbink *et al*., 1998; Salminen *et al*., 1996; Vetriani *et al*., 1998; Yip *et al*., 1995; Sindelar *et al*., 1998). All these salt bridges can be broadly devided into two main classes:

(i) <u>Intra-helical or local:</u> This class arises out of side chain interaction between the charged residues in a helix. Biophysical studies have specifically revealed that intra-helical EK, ER, DK, DR salt bridge pairs with the separation of 3 and 4 stabilize helices (Huyghues-Despointes *et al*., 1993; Scholtz *et al*., 1993).

(ii) Tertiary: This class occurs as a result of interaction between non-local charged residues of proteins. The class includes a large variety of salt bridges, such as inter-helical salt bridges, helix-sheet salt bridges and intersubunit salt bridges.

Figure 1A illustrates local and tertiary salt bridges. We studied the contribution of both types of salt bridges to protein thermostability. In addition to salt bridges discussed, other electrostatic interactions, such as charge-helix dipole interaction, can also stabilize proteins. Previous research had shown that negatively charged residue can interact with the positive side of the helix dipole and thereby stabilize thermophilic proteins (Nicholson *et al*., 1991; Aqvist *et al*., 1991; Tidor *et al*., 1991). In our study we also analyzed the results to see whether such interactions are important for protein thermostability.

A Genome-wide Comprehensive Study: Our Goals and Strategy

Most of the studies about protein thermostability referred to above involved analysis of only a few proteins. With the advent of fast DNA sequencing technology, complete genome sequences of several organisms are now available (Devine *et al*., 1995). As a result, it is now possible to comprehensively study all the proteins in an organism and compare the results with those of other organisms. We have done a number of similar analyses, comparing various aspects of protein structure, such as secondary structural composition and fold usage, between several recently sequenced genomes (Gerstein, 1997, 1998a,b; Hegyi & Gerstein *et al*., 1999). Similar kind of studies also has been carried out by other investigators (Frishman & Mewes, 1997; Sanchez & Sali, 1998; Wolf *et al*., 1999; Kyrpides & Ouzounis, 1999; Fetrow *et al*., 1998; Amano *et al*, 1997).

In this investigation our aim is to study the effect of salt bridges, as well as other stability factors such as deamidation and protein length, in thermophiles and mesophiles by comprehensive analysis of all protein sequences in their genomes. All the factors that are studied here are shown in the Figure 1A. We have analyzed the genomes of twelve organisms, of which four archaeons and one hyper-thermophilic eubacteria are grouped together as thermophiles. The rest, one eukarya and six eubacteria, are grouped together as mesophiles. These organisms are listed in Table 1.

Figure 1A: Factors Studied.

Table 1: List of Organisms.

Our overall strategy of salt bridge analysis is shown in Figure 1B-D. In the first step of our study we calculated the amino acid composition of all twelve genomes.

Figure 1B, 1C, 1D: Overall Strategy.

In the second step we focused on the intra-helical salt bridges, calculating the frequency of putative salt bridge pairs in the helices of proteins, and analyzed the data. Since it is possible that the frequency of intra-helical salt bridges in a genome may be biased due to sequence repeats and multiple paralogs specific to the organism, we analyzed a small set of orthologous proteins that are present in all the organisms and performed a similar calculation on this set.

In the third step we looked into the frequency of tertiary salt bridges in proteins in all the genomes. Within the above orthologous set, we calculated this frequency in only those proteins whose structures are known.

RESULTS AND DISCUSSION

Study of Local Salt Bridges

Amino Acid Composition in the Genome, overall and just in Helices

We estimated amino acid composition both in the entire genome and protein helices. These are shown in Figures 2A and 2B.

Since secondary structures of most of the proteins in genomes are unknown, we predicted the secondary structures of proteins to calculate the amino acid composition in helices. Secondary structure prediction of the proteins was performed using GOR(IV) program. (Garnier *et al*., 1996, 1978; Gibrat *et al*., 1987). This is a well-established and commonly used method. It is statistically based, so that the prediction for a particular residue to be in a given state (say, Ala to be in a helix) is based directly on the frequency of the residue's occurrence in that state in a database of solved structures (taking into account neighbors at ±1, ±2 and so forth). The GOR method uses only single sequence information compared to current 'state-of-the-art' methods that incorporate multiple sequence information (King *et al*., 1996; Rost *et al*., 1996a,b). While single sequence predictions are slightly less accurate than multiple sequence methods (65% versus 71%), we felt that using single sequence methods avoids various bias problems that can plague multiple sequence methods – i.e. we can only get multiple sequence information for a biased sample from each genome. Furthermore, we felt that the difference in accuracy between single and multiple sequence methods was not so vital in the overall context of our study, given our focus on bulk-averaged results.

Figure 2A: Genome composition.

Figure 2B: Helix composition.

It is observed that both at the overall genomic level and in the helices, the amounts of glutamate, lysine and arginine (E, K, R) are higher in thermophilic proteins than in mesophilic proteins. This increase in charged residues suggests that in general we can expect to see more salt bridges in thermophiles than in mesophiles. Analysis of amino acid composition has shown that the amount of negatively charged aspartate residue remained almost the same in all the organisms.

Use of Log of Odds (LOD) Calculation: Abundance of Local Salt bridges

As a result of the high content of charged residues in the helices we can generally expect to see a greater number of intra-helical salt bridges in thermophiles than in mesophiles. In order to see whether salt bridges are even more numerous than this elevated *a priori* "baseline", we calculated an odds ratio for all possible 400 amino acid pairs in helices. As described more fully in the caption to Figure 1B, this is essentially the observed number of occurrences of a given pair, divided by its expected number if there was no correlation -- e.g. the frequency of EK(3) divided by product of the individual E and K frequencies. Note that here the notation EK(3) implies the EK salt bridge pair with a separation of 3; we used similar notations throughout the text. We then took the logarithm of this odds ratio, arriving at a log odds (LOD) value. LOD values represent a measure of relative abundance for each pair in helices. Therefore, a higher LOD value for a particular pair would mean a higher frequency of that pair than other pairs in the genome. We calculated the LOD values for 400 amino acid pairs with a separation from 1 to 6. It is observed that for any salt bridge pair, its LOD values peak at the separation of either 3 or 4, indicating that these pairs probably represent intra-helical salt bridges, as suggested by the previous biophysical studies. As an illustration of this general result we plotted the LOD values of EK pairs at various separations in helices in Figure 3. Note that the LOD values of EK pairs peak at the separation of 3 for all the organisms.

Figure 3: LOD values of EK pairs in helix.

Results of LOD value calculation show that the LOD values for salt bridge pairs EK, ER and DR with the separation of 3 and 4 are generally higher in helices of thermophilic proteins than in mesophilic ones. In order to see whether the charged residues in the strand part of the protein sequences were correlated in a salt-bridge fashion, we performed a similar LOD-value calculation on the strands and compared the result with that for helices. Similarly, we calculated genome-wide LOD values for the pairs by performing calculations on entire protein sequences. Comparison of the results shows that the LOD values for salt-bridge pairs are higher in helices than in other secondary structural elements and that this is true to a greater degree in thermophilic organisms. Our results thus imply that the charged amino acid residues are not only more numerous in thermophiles than mesophiles, but are also more highly correlated in helices of thermophilic proteins with a salt-bridge separation of 3 and 4.

Table 2A, 2B: LOD values of salt bridge pairs.

Correlation between Temperature and Salt Bridge Frequency

We computed the LOD values for EK pairs with the separation of 3 and 4, and the result is shown in Figure 4A. The figure shows that the LOD values are higher for thermophiles than for mesophiles. From the figure it is also observed that in the thermophilic region, LOD values increase from MT to OT commensurate with the steady increase in physiological temperatures from MT (65°C) to OT (98°C). This correlation of physiological temperatures with the intra-helical salt bridge frequency suggests that higher temperatures require a greater number of salt bridges to stabilize the helices in proteins.

Figure 4A: LOD values of EK pairs with the separations of 3 and 4.

Helix Dipole Stabilization

In our LOD calculation in helices we found that the values for EK(3) and EK(4) pairs are always higher than the corresponding values for KE pairs (data not shown). This variation of LOD values on the orientation of the charged pair is significant in terms of charge-helix dipole interaction. Since negatively charged glutamate residue can stabilize a helix by interacting with the positive amino end of a helix dipole (Nicholson *et al*., 1991; Eijsink *et al*., 1992; Aqvist *et al*., 1991; Tidor & Karplus, 1991), this observation indicates that the thermophilic proteins gain stability from helix dipole stabilization.

Analyses of Control for Biases in the Statistics

The Problem of Bias in Our Comprehensive Genome-wide Statistics

While doing genome-wide surveys, one has to be careful to assess the degree to which one's calculated statistics could be biased. With regard to this, there are a number of specific issues relevant here. Firstly, sequence repeats, e.g. repetitive charged sequences in a set of thermophilic proteins, could skew the results. Secondly, unique protein sequences enriched in salt bridges, could be highly duplicated in the thermophile genomes (forming large paralogous families), and this could also influence our results (see, for instance, figure 1C). A similar situation may arise involving only the sequences unique to mesophiles. We therefore need to test the significance of LOD results and verify our conclusions with statistical controls and alternate procedures.

Rank Statistics

One technique to test the significance of our results is the use of rank statistics. Here the idea is that if we arrange the LOD values of all 400 pairs for each separation in an ordered list and observe that a particular pair --- EK(3), for example --- is at the top of the list, then we could infer that this pair is among the most over-represented in the helices of the proteins for that organism. Table 3 summarizes the rank statistics for salt-bridge pairs that ranked in the top 20 of a possible 400. The results show that while the ranks of salt-bridge pairs vary greatly among all twelve genomes, the ranks of EK(3) pairs are generally higher for thermophiles compared to mesophiles in helices. MT is an exception to this general trend. In contrast, when the non-helical regions are considered, this distinction lessens.

Table 3: Rank statistics.

Random Resampling

We directly addressed the problem of sequence repeats by a random resampling procedure. We simulated thermophilic and mesophilic genomes by randomly drawing proteins from two large pools of thermophilic and mesophilic sequences. From these simulated genomes we calculated the LOD values for charged amino acid pairs in helices. Figure 4B shows the distribution of these values for the EK(3) pair. Note the distinct difference in the distributions. Statistical tests were performed to estimate the degree of significance of this difference, and it was found that given the width of the distribution, the chance that any mesophile could have a LOD value similar to a thermophile is less than 5% (for EK(3) and EK(4)). This implies that our LOD calculation results are significant in a statistical sense. (These calculations are described in more detail in the figure caption.)

Figure 4B: Distribution of LOD values of EK(3) for randomly generated mesophilic and thermophilic genomes.

Stratified Resampling using Orthologs

Another way of removing biases is through the use of stratified sampling procedures (Anderson & Finn, 1996). The idea here can most easily be described in terms of a demographic comparison of a particular characteristic between populations -- for example, height in northern versus southern populations. It is possible that the overall population could be fractionated into further subdivisions on another parameter, potentially linked to height, say age (old vs. young). Our initial analysis above of salt bridge statistics was analogous to computing the average height over the entire population irrespective of age. However, the possibility that one population has more of a certain age group than another could potentially skew the statistics (e.g. Northerners are older and taller). To compensate for such bias in the sample we could take a representative sample from every age group and calculate the average height for that strata. This is what we did in stratified sampling to study the salt bridge abundance.

Our strata were sets of orthologous proteins present in each of the 12 genomes. Orthologous proteins evolved from a common ancestral gene and usually share the same structure and function (Fitch, 1970). Statistics obtained from sets of orthologous proteins can be considered to be relatively free from bias arising from sequence repeats or large paralogous families. In our study we selected 52 sets of orthologous proteins (listed on our website). Our ortholog selection strategy is explained in detail in Figure 1C. It was derived using the cluster-of-orthologous groups (COGs) approach (Tatusov *et. al*., 1997). We used only COGs for which we could determine a single best representative for each genome, and we extended the initial COGs assignments (currently 8 genomes) to include all twelve genomes in our study.

On our set of 52 orthologous proteins we performed analyses similar to what we did on the entire genome. Composition analysis showed a similar trend of increasing amounts of charged residues from mesophile to thermophile, as was observed in the overall genome analysis (Figure 2C). Note in the figure that the hyperthermophilic eubacteria *Aquifex aeolicus* has moved closer in position to the other eubacteria, perhaps indicating that some exclusively archaeal paralogous family is heavily weighted with charged residues. Likewise, we calculated LOD values for our set of 52 orthologs. The results for important salt bridge pairs are shown in Table 2C. Although the LOD values for EK(3) had decreased for both thermophiles and mesophiles, thermophiles still maintained higher average LOD values for EK(3), EK(4), DR(3) and ER(4). This result is important: in spite of involving only 52 groups of proteins, the stratified resampling comparisons showed putative salt bridge frequency was clearly higher in thermophiles than mesophiles.

Figure 2C: Amino acid composition of 52 COG proteins in helices.

Table 2C: LOD values of ion pairs for 52 COG proteins.

Study of Tertiary Salt Bridges

So far, our study of salt bridges has focussed only on intra-helical salt bridges. Moreover, these statistics depend on accuracy in the prediction of protein secondary structures. Therefore, to complement our conclusions on intra-helical salt bridge abundance, we studied the tertiary salt bridges in thermophilic and mesophilic proteins of known structure. Here, we followed a procedure similar to that performed by Schueler & Margalit (1995). Since any such study of tertiary salt-bridges requires the knowledge of detailed protein 3D structure, which is unknown for most proteins in the genome, we tried the following strategy, schematized in Figure 1D. Where possible, we mapped the sequence of a protein with a known 3D structure onto a corresponding orthologous group of sequences to identify the putative tertiary salt bridges in the new sequences. This approach rests on the idea that since orthologous proteins conserve their structures, knowledge of one protein structure can be extended to others in the same group. More specifically, we took query sequences from each of our 52 orthologous groups of proteins and compared them with the PDB structural database by pairwise sequence comparison (Lipman & Pearson, 1985; Sussman *et. al*., 1998). This resulted in a list of 18 PDB structures that map onto corresponding orthologous groups. As listed in Table 4, we classified these 18 orthologous groups of known structure into three categories: (i)

Ribosomal proteins, (ii) Amino-acyl tRNA synthetases, and (iii) Other proteins (including proteins with various functions).

Using the strategy outlined in Figure 1D, we obtained rough estimates of the number of salt bridges for each protein in the 18 orthologous groups of known structure. Table 4 shows some summary statistics based on these numbers. It shows that for two categories, ribosomal proteins and tRNA-synthetases, thermophiles have somewhat more tertiary salt bridges than mesophiles. For proteins in the "other" category, there is less of a difference between thermophiles and mesophiles.

Table 4: Statistics with raw salt bridge counts for 18 orthologous groups.

Other Stabilizing Factors

So far we have discussed the role of salt-bridge interactions in thermophilic proteins. One should note here that since structures for most of the proteins are unknown, it was not possible for us to study the contribution of other factors, such as the effect of hydrophobic internal packing on protein thermal stability. However, in addition to salt-bridges, we also studied the effect of two other factors on thermostability of proteins and compared their results with that of the salt-bridges: (i) deamidation (ii) protein length.

Deamidation

Studies showed that glutamine and asparagine undergo a deamidation reaction leading to instability in a protein (Catanzano *et al*., 1997). Therefore reduction in the amount of these two amino acids can stabilize proteins. In our amino acid composition study, we observed that compared to mesophiles, the amounts of glutamine (N) and asparagine (Q) have decreased in thermophiles. Furthermore, we noticed that among hydrophobic amino acids the amount of valine and isoleucine have increased in thermophiles. In this context it is to be noted that amount of proline which is believed to contribute thermal stability in proteins (Matthhews *et al.*, 1987; Hardy *et al,* 1993; Wallon *et al.*, 1997) did not exhibit any bias and remained almost same both in thermophiles and mesophiles.

Protein Length and Thermal Stability: The Contradictory Position of Aquifex aeolicus

It has been argued that shorter protein length increases the compactness of the protein and reduces its flexibility. A biophysical study by Nagi & Regan (1997) had suggested that there is an inverse correlation between loop length and protein stability. Thompson & Eisenberg (1999) in a recent study put forward a thermodynamic argument supporting this correlation, and showed that the thermophilic proteins have a higher tendency towards shorter loops than its mesophilic counterparts, by comparing homologous proteins from the genomes of a large number of organisms. In our study, we analyzed the sequence length distribution of proteins for all organisms to understand how protein length is related to thermostability. Our results, shown in Figure 5A, indicate that the length distributions for thermophiles do indeed fall off more rapidly than the ones for mesophiles. Furthermore, when we fit curves to the length distribution of just thermophilic or just mesophilic proteins, we found that the median (and mode) length was less in the thermophiles than in the mesophiles (Figure 5B). This result was true for both the genome overall and for just our sample of 52 orthologs. Therefore, our "first pass" results on protein length appear to support the notion that the proteins in thermophiles are shorter than those in mesophiles.

Figure 5A: Length distribution of twelve organisms.

Figure 5B: Length distribution in terms of protein content.

However, when we looked at the sequence lengths in further detail, we found that the story was more complicated. The distribution of protein lengths for *Aquifex aeolicus*, a hyperthermophilic eubacteria, is more similar to those of the other mesophilic eubacteria than to the other thermophilic organisms, which are all archaeas. Furthermore, yeast appears to have distinctly longer proteins than those in either of the prokaryotic kingdoms. The distribution of protein lengths therefore appears to be more related to kingdom than to environment -- reflecting historical contingency rather than chemical necessity. This result is illustrated in Figure 6A, which shows how "phylogenetic composition" of proteins with a given length becomes progressively less archaeal and more eukaryotic as one moves to longer proteins. This result is further borne out in the table of Figure 6B, where it can be seen how average protein lengths are correlated with kingdom. In this table we included average protein length for *C. elegans*, the other known eukaryotic genome, to illustrate that the long sequences are characteristic of other eukaryotes beside yeast.

Figure 6A: Length distribution as %content contribution.

Figure 6B: Average length of protein in twelve organisms.

CONCLUSION

From the comparison of our results on amino-acid composition and LOD statistics, we argue that the occurrence of excess intra-helical salt bridges in thermophiles may have its origin in two factors. Firstly, the thermophiles have a higher content of charged amino acids than the mesophiles. Secondly, these charged residues are more preferentially arranged with a 1,4 salt-bridge spacing in thermophilic helices than in mesophilic ones. Since the results of our calculations on orthologous groups of proteins were similar to our genome-wide results, we infer that the sequence repeats or paralogous sequence families do not skew the observed abundance of intra-helical salt bridges in thermophiles. Our results also showed that the thermophilic proteins have higher occurrence of tertiary salt bridges than the mesophilic proteins. Thus we conclude that the salt-bridge interactions play a vital role in stabilizing thermophilic proteins. Our study also showed that, in addition to salt-bridges, there are other factors that can contribute to protein thermal stability. Reduction of deamidation by decreasing the amounts of glutamine and asparagine in proteins confers stability to thermophilic proteins. Though we examined the contribution of protein sequence lengths, we found that they are only loosely connected with the protein thermostability. Therefore, among all the three factors that we studied here, we found that while the extent of contribution to thermostability varies for each factor, salt-bridge contribution is most consistent with the increasing physiological temperatures and one of the most important factors for protein thermostability.

REFERENCE

Amano, N., Ohfuku, Y. & Suzuki, M. (1997). Genomes and DNA conformation. *Biol. Chem.* **378**(12), 1397-404.

Anderson, T. W. & Finn, J. D. (1996). *The New Statistical Analysis of Data*. **17**, 644 Springer Verlag, New York, Berlin.

Aqvist, J., Luecke, H., Quiocho, F. A. & Warshel, A. (1991). Dipoles localized at helix termini of proteins stabilize charges. *Proc. Natl. Acad. Sci. USA,* **88**(5), 2026-30.

Auerbach, G., Ostendrop, R., Prade, L., Korndorfer, I., Dams, T., Huber, R., Jaenicke, R. (1998). Lactate dehydrogenase from the hyperthermophilic bacterium thermotoga maritima: the crystal structure at 2.1 A resolution reveals strategies for intrinsic protein stabilization. *Structure* **8**, 769-81.

Blattner, F. R., Plunkett, G., 3rd, Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., Gregor, J., Davis, N. W., Kirkpatrick, H. A., Goeden, M. A., Rose, D. J., Mau, B. & Shao, Y. (1997). The complete genome sequence of Escherichia coli K-12. *Science* **277**(5331), 1453-74.

Bult, C. J., White, O., Olsen, G. J., Zhou, L., Fleischmann, R. D., Sutton, G. G., Blake, J. A., FitzGerald, L. M., Clayton, R. A., Gocayne, J. D., Kerlavage, A. R., Dougherty, B. A., Tomb, J. F., Adams, M. D., Reich, C. I., Overbeek, R., Kirkness, E. F., Weinstock, K. G., Merrick, J. M., Glodek, A., Scott, J. L., Geoghagen, N. S. M. & Venter, J. C. (1996). Complete genome sequence of the methanogenic archaeon, Methanococcus jannaschii. *Science* **273**(5278), 1058-73.

Catanzano, F., Graziano, G., Capasso, S. & Barone, G. (1997). Thermodynamic analysis of the effect of selective monodeamidation at asparagine 67 in ribonuclease A. *Protein Sci.* **6**(8), 1682-93.

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

Colacino, F., Crichton, R. R. Enzyme thermostabilization: the state of the art. (1997). Biotechnol. Genet. Eng. Rev.**14**, 211-77.

Deckert, G., Warren, P. V., Gaasterland, T., Young, W. G., Lenox, A. L., Graham, D. E., Overbeek, R., Snead, M. A., Keller, M., Aujay, M., Huber, R., Feldman, R. A., Short, J. M., Olsen, G. J. & Swanson, R. V. (1998). The complete genome of the hyperthermophilic bacterium Aquifex aeolicus. *Nature* **392**(6674), 353-8.

Devine, K. M. & Wolfe, K. (1995). Bacterial genomes: a TIGR in the tank. *Trends Gen.* **11**(11), 429-31.

Eijsink, V. G., Vriend, G., Van der Zee, J. R., Van den Burg, B., Venema, G. (1992) Increasing the thermostability of the neutral proteinase of Bacillus stearothermophilus by improvement of internal hydrogen-bonding. *Biochem. J.* **285**(Pt 2), 625-8.

Elcock, A. H. & McCammon, J. A. (1997). Continnum Solvation Model for Studying Protein Hydration Thermodynamics at High Temperature. *J. Phys. Chem.* **B**(101), 9624-34.

Fetrow, J. S., Godzik, A. & Skolnick, J. (1998). Functional analysis of the Escherichia coli genome using the sequence-to-structure-to-function paradigm: identification of proteins exhibiting the glutaredoxin/thioredoxin disulfide oxidoreductase activity. *J. Mol. Biol.* **282**(4), 703-11

Fitch, W. M. (1970). *Syst. Zool.* **19**, 99.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A. & Merrick, J. M. (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* **269**(5223), 496-512.

Frishman, D. & Mewes, H. W. (1997). Protein structural classes in five complete genomes. *Nature Struct. Biol.* **4**, 626-628.

Garnier, J., Gibrat, J. F. & Robson, B. (1996). GOR method for predicting protein secondary structure from amino acid sequence. *Methods Enzymol.* **266**, 540-53.

Garnier, J., Osguthorpe, D. J. & Robson, B. (1978). Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.* **120**(1), 97-120.

Gerstein, M. (1997). A Structural Census of Genomes: Comparing Eukaryotic, Bacterial and Archaeal Genomes in terms of Protein Structure. *J. Mol. Biol.* **274**, 562-576.

Gerstein, M. (1998a). How Representative are the Known Structures of the Proteins in a Complete Genome? A Comprehensive Structural Census. *Folding Design* **3**, 497-512.

Gerstein, M. (1998b). Patterns of Protein-Fold Usage in Eight Microbial Genomes: A Comprehensive Structural Census. *Proteins* **33**, 518-534.

Gibrat, J. F., Garnier, J. & Robson, B. (1987). Further developments of protein secondary structure prediction using information theory. New parameters and consideration of residue pairs. *J. Mol. Biol.* **198**(3), 425-43.

Goffeau *et al*.,. (1997). The yeast genome directory. *Nature* **387**(6632 Suppl), 5.

Gupta, M. (1995). *Thermostability of Enzymes*, Springer, Berlin.

Hardy, F., Vriend, G., Veltman, O. R., van der Vinne, B., Venema, G., Eijsink, V. G. H. (1993). Stabilization of *Bacillus stearothermophilus* neutral protease by introduction of of prolines. *FEBS Lett.* **317**(1,2), 89-92.

Hegyi, H. & Gerstein, M. (1999). The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J. Mol. Biol.* **288**(1), 147-64.

Hennig, M., Darimont, B., Sterner, R., Kirschner, K. & Jansonius, J. N. (1995). 2.0 A structure of indole-3-glycerol phosphate synthase from the hyperthermophile Sulfolobus solfataricus: possible determinants of protein stability. *Structure* **3**(12), 1295-306.

Hennig, M., Sterner, R., Kirschner, K. & Jansonius, J. N. (1997). Crystal structure at 2.0 A resolution of phosphoribosyl anthranilate isomerase from the hyperthermophile Thermotoga maritima: possible determinants of protein stability. *Biochemistry* **36**(20), 6009-16.

Higgins, D. G., Thompson, J. D. & Gibson, T. J. (1996). Using CLUSTAL for multiple sequence alignments. *Methods in Enzymology* **266**, 383-402.

Himmelreich, R., Hilbert, H., Plagens, H., Pirkl, E., Li, B. C. & Herrmann, R. (1996). Complete sequence analysis of the genome of the bacterium Mycoplasma pneumoniae. *Nucl. Acids Res.* **24**(22), 4420-49.

Huyghues-Despointes, B. M., Scholtz, J. M. & Baldwin, R. L. (1993). Helical peptides with three pairs of Asp-Arg and Glu-Arg residues in different orientations and spacings. *Protein Sci.* **2**(1), 80-5.

Jaenicke, R. & Bohm, G. (1998). The stability of proteins in extreme environments. *Curr. Opin. Struct. Biol.* **8**(6), 738-48.

Kaneko, T., Sato, S., Kotani, H., Tanaka, A., Asamizu, E., Nakamura, Y., Miyajima, N., Hirosawa, M., Sugiura, M., Sasamoto, S., Kimura, T., Hosouchi, T., Matsuno, A., Muraki, A., Nakazaki, N., Naruo, K., Okumura, S., Shimpo, S., Takeuchi, C., Wada, T., Watanabe, A., Yamada, M., Yasuda, M. & Tabata, S. (1996). Sequence analysis of the genome of the unicellular cyanobacterium Synechocystis sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res.* **3**(3), 109-36.

Kawarabayasi, Y., Sawada, M., Horikawa, H., Haikawa, Y., Hino, Y., Yamamoto, S., Sekine, M., Baba, S., Kosugi, H., Hosoyama, A., Nagai, Y., Sakai, M., Ogura, K., Otsuka, R., Nakazawa, H., Takamiya, M., Ohfuku, Y., Funahashi, T., Tanaka, T., Kudoh, Y., Yamazaki, J., Kushida, N., Oguchi, A., Aoki, K. & Kikuchi, H. (1998). Complete sequence and gene organization of the genome of a hyper-thermophilic archaebacterium, Pyrococcus horikoshii OT3. *DNA Res.* **5**(2), 55-76.

King, R. D. & Sternberg, M. J. E. (1996). Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. *Protein Sci.* **5**, 2298-2310.

Klenk, H. P., Clayton, R. A., Tomb, J. F., White, O., Nelson, K. E., Ketchum, K. A., Dodson, R. J., Gwinn, M., Hickey, E. K., Peterson, J. D., Richardson, D. L., Kerlavage, A. R., Graham, D. E., Kyrpides, N. C., Fleischmann, R. D., Quackenbush, J., Lee, N. H., Sutton, G. G., Gill, S., Kirkness, E. F., Dougherty, B. A., McKenney, K., Adams, M. D., Loftus, B. & Venter, J. C. (1997). The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon Archaeoglobus fulgidus. *Nature* **390**(6658), 364-70.

Knapp, S., de Vos, W. M., Rice, D. & Ladenstein, R. (1997). Crystal structure of glutamate dehydrogenase from the hyperthermophilic eubacterium Thermotoga maritima at 3.0 A resolution. *J. Mol. Biol.* **267**(4), 916-32.

Koonin, E. V. & Galperin, M. Y. (1997). Prokaryotic genomes: the emerging paradigm of genome-based microbiology. *Curr Opin Genet Dev* **7**(6), 757-63.

Korndorfer, 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 A resolution. *J. Mol. Biol.* **246**(4), 511-21.

Kyrpides, N. C. & Ouzounis, C. A. (1999). Transcription in Archaea. *Proc. Natl. Acad. Sci. USA*, **96**(15), 8545-50

Lebbink, J. H., Knapp, S., van der Oost, J., Rice, D., Ladenstein, R. & de Vos, W. M. (1998). Engineering activity and stability of Thermotoga maritima glutamate dehydrogenase. I. Introduction of a six-residue ion-pair network in the hinge region. *J. Mol. Biol.* **280**(2), 287-96.

Lipman, D. J. & Pearson, W. R. (1985). Rapid and sensitive protein similarity searches. *Science* **227**(4693), 1435-41.

Nagi, A. D. & Regan, L. (1997). An inverse correlation between loop length and stability in a four-helix-bundle protein. *Folding Design* **2**(1), 67-75.

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

Nicholson, H., Anderson, D. E, Dao-pin, S., Matthews, B. W. (1991). Analysis of the interaction between charged side chains and the α-helix dipole using designed thermostable mutants of phage T4 lysozyme. *Biochemistry* **30,** 9816-28.

Perutz, M. F., Raidt, H. (1975). Stereochemical basis of heatstability in bacterial ferredoxins and in haemoglobin A2. *Nature* **255**(5505), 256-9.

Petukhov, M., Kil, Y., Kuramitsu, S., Lanzov, V. (1997), Insights into thermal resistance of proteins from intrinsic stability of their α-helices. *Proteins*, **29**, 309-20.

Querol, E., Perez-Pons, J. A. & Mozo-Villarias, A. (1996). Analysis of protein conformational characteristics related to thermostability. *Protein Eng.* **9**(3), 265-71.

Rost, B. (1996a). PHD: predicting one-dimensional protein structure by profile-based neural networks. *Methods Enzymol.* **266**, 525-39.

Rost, B., Fariselli, P. & Casadio, R. (1996b). Topology prediction for helical transmembrane segments at 95% accuracy. *Prot. Sci.* **7**, 1704-1718.

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 A resolution. *Biochemistry* **36**(33), 9983-94.

Russell, R. J. M. & Taylor, G. L. (1995). Engineering thermostabilty: lessons from thermophilic proteins. *Curr. Opin. Biotechnol.* **6**, 370-74.

Sanchez, R. & Sali, A. (1998). Large-scale protein structure modeling of the Saccharomyces cerevisiae genome. *Proc. Natl. Acad. Sci. USA,* **95**(23), 13597-602.

Sindelair, C. V., Hendsch, Z. S. & Tidor, B. (1998). Effects of salt bridges on protein structure and design. *Protein Science* **7**(9), 1898-914.

Salminen, T., Teplyakov, A., Kankare, J., Cooperman, B. S., Lahti, R. & Goldman, A. (1996). An unusual route to thermostability disclosed by the comparison of Thermus thermophilus and Escherichia coli inorganic pyrophosphatases. *Protein Sci.* **5**(6), 1014-25.

Scholtz, J. M., Qian, H., Robbins, V. H. & Baldwin, R. L. (1993). The energetics of ion-pair and hydrogen-bonding interactions in a helical peptide. *Biochemistry* **32**(37), 9668-76.

Schueler, O. & Margalit, H. (1995). Conservation of salt bridges in protein families. *J. Mol. Biol.* **248**(1), 125-35.

Scandurra, R., Consalvi, V., Chiaraluce, R., Politi, L., Engel, P. C. (1998). Protein thermostability in extremophiles. *Biochimie* **80**(11), 933-41.

Smith, D. R., Doucette-Stamm, L. A., Deloughery, C., Lee, H., Dubois, J., Aldredge, T., Bashirzadeh, R., Blakely, D., Cook, R., Gilbert, K., Harrison, D., Hoang, L., Keagle, P., Lumm, W., Pothier, B., Qiu, D., Spadafora, R., Vicaire, R., Wang, Y., Wierzbowski, J., Gibson, R., Jiwani, N., Caruso, A., Bush, D. & Reeve, J. N. (1997). Complete genome sequence of Methanobacterium thermoautotrophicum deltaH: functional analysis and comparative genomics. *J. Bacteriol.* **179**(22), 7135-55.

Suckow, J. M., Amano, N., Ohfuku, Y., Kakinuma, J., Koike, H. & Suzuki, M. (1998). A transcription frame-based analysis of the genomic DNA sequence of a hyper-thermophilic archaeon for the identification of genes, pseudo-genes and operon structures. *FEBS Lett.* **426**(1), 86-92.

Sussman, J. L., Lin, D., Jiang, J., Manning, N. O., Prilusky, J., Ritter, O. & Abola, E. E. (1998). Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules. *Acta Cryst. Section D-Biol. Cryst.* **54**(1 (Pt 6)), 1078-84.

Szilagyi, A., Zavodszky, P. (1995). Structural basis for the extreme thermostability of D-glyceraldehyde-3-phosphate dehydrogenase from Thermotoga maritima: analysis based on homology modelling. *Protein Eng.* **8**(8), 779-89.

Tatusov, R. L., Koonin, E. V. & Lipman, D. J. (1997). A genomic perspective on protein families. *Science* **278**(5338), 631-7.

Thompson, M. J. & Eisenberg, D. (1999). Transproteomic Evidence of a Loop-Deletion Mechanism for Enhancing Protein Thermostability. *J. Mol. Biol.* **290**(2), 595-604.

Tidor, B. & Karplus, M. (1991). Simulation analysis of the stability mutant R96H of T4 lysozyme. *Biochemistry* **30**(13), 3217-28.

Tomb, J. F., White, O., Kerlavage, A. R., Clayton, R. A., Sutton, G. G., Fleischmann, R. D., Ketchum, K. A., Klenk, H. P., Gill, S., Dougherty, B. A., Nelson, K., Quackenbush, J., Zhou, L., Kirkness, E. F., Peterson, S., Loftus, B., Richardson, D., Dodson, R., Khalak, H. G., Glodek, A., McKenney, K., Fitzegerald, L. M., Lee, N., Adams, M. D. & Venter, J. C. (1997). The complete genome sequence of the gastric pathogen Helicobacter. *Nature* **388**(6642), 539-47.

Vetriani, C., Maeder, D. L., Tolliday, N., Yip, K. S., Stillman, T. J., Britton, K. L., Rice, D. W., Klump, H. H. & Robb, F. T. (1998). Protein thermostability above 100 degreesC: a key role for ionic interactions. *Proc. Natl. Acad. Sci. USA*, **95**(21), 12300-5.

Vieille, C., Burdette, D. S. & Zeikus, J. G. (1996). Thermozymes. *Biotechnol. Annu. Rev.* **2**, 1-83.

Vieille, C. & Zeikus, J. G. (1996). Thermozymes: identifying molecular determinants of protein structural and functional stability. *Trends Biotechnol.* **14**, 183-90.

Vogt, G. & Argos, P. (1997a). Protein thermal stability: hydrogen bonds or internal packing? *Folding Design*, **2**(4), S40-6.

Vogt, G., Woell, S. & Argos, P. (1997b). Protein thermal stability, hydrogen bonds, and ion pairs. *J. Mol. Biol.* **269**(4), 631-43.

Wallon, G., Kryger, G., Lovett, S. T., Oshima, T., Ringe, D., Petsko, G. A. (1997). Crystal structures of *Escherichia coli* and *Salmonella typhimurium* 3-lsopropylmalate dehydrogenase and comparison with their thermophilic counterpart from *Thermus thermophilus*. *J. Mol. Biol*. **266**, 1016-31.

Wolf, Y. I., Brenner, S. E., Bash, P. A. & Koonin, E. V. (1999). Distribution of protein folds in the three superkingdoms of life. *Gen. Res.* 9(1), 17-26.

Yip, K. S., Stillman, T. J., Britton, K. L., Artymiuk, P. J., Baker, P. J., Sedelnikova, S. E., Engel, P. C., Pasquo, A., Chiaraluce, R. & 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**(11), 1147-58.

Xiao, L. & Honig, B. (1999). Electrostatic contribution to the stability of hyperthermophilic proteins. *J. Mol. Biol.* **5,** 1435-44.

[TABLE CAPTIONS]

Table 1. List of Organisms.

The table lists the twelve organisms whose sequences are used in calculation. Column three shows the two-letter abbreviations for the genomes of the organisms listed in the first column. The fourth column lists the number of open reading frames found in the genome. The last column shows the physiological temperatures of thermophiles. For mesophiles we referred to 'mesophilic temperatures' which range from 10 to 45 °C. Data-files of predicted proteins were taken from the websites referred to in the papers above, with the exception of OT for which predicted proteins were from the analysis of Suckow *et al*. (1998).

Table 2: LOD values (values are in percentages).

Table 2A and 2B show LOD values of salt bridge pairs in helix and genome. Since in helices salt bridge pairs at the separation of 3 and 4 are known to stabilize proteins, we have listed their LOD values separately. LOD values of the salt bridge pairs in strands are not shown here, as it is obvious from the whole genome results. Table 2C lists the LOD values of the ion pairs for 52 orthologous proteins. Note that LOD values for the slat bridge pairs remained high even in the small set of 52 orthologous proteins.

Table 3. Rank statistics of salt bridge pairs.
The ranks of other salt bridge pairs (ER(3), EK(4) and DR(3)) were not remarkably different between thermophiles and mesophiles. A similar study on the predicted strand sequence did not show any significant ranking for salt bridge pairs (results are not shown).

Table 4. Statistics with tertiary salt bridges.

This table summarizes the results of tertiary salt bridge counts. Column one shows the COG identifiers for the orthologous groups that are selected. Second column gives the functional class for the each of this group and the fourth column lists the PDB identifiers for homologous proteins with known structures. Third column represents our category. For every protein, we calculated the average number of salt bridges present in thermophiles and in mesophiles as shown in columns 5 and 6. Column 7 shows the difference between the two. Based on this difference we set up a scoring scheme that qualitatively describes the relative abundance of tertiary salt bridges. If the difference is > 1.0, a positive (+) sign is assigned showing a predominance of salt bridges in thermophiles; if the difference is < -1.0, a negative (-) sign is assigned showing a predominance of salt bridges in mesophiles; for any other value of difference, no sign is assigned to either thermophiles or mesophiles, thus showing no bias for salt bridges. Note that in the two main categories (ribosomal proteins and tRNA synthetases thermophiles) thermophiles have a higher amount of tertiary salt bridges than mesophiles.

[FIGURE CAPTIONS]

[CAPTION]

Figure 1A. Three factors in a protein that are studied for their contribution to protein thermostability.

(i) Local salt bridges (ii) tertiary salt bridges (iii) protein length.

Figure 1B. Determining the position of local (intra-helical) and tertiary salt bridges.

The box in the figure represents a protein sequence with known structure and each ± combination connected with dotted lines represents a salt bridge pair as observed in the structure. Since the EK pair involves an interaction between the charged residues in two separate secondary structural elements, it is defined as a tertiary salt bridge. Similarly the DK pair occurring within a helix is termed a local salt bridge. Using this definition we calculated the LOD values for the intra-helical amino acid pair as follows. The odds ratio R for any particular pair, say XY, at a separation i is defined by,

$$R[XY(i)] = \frac{\text{Observed number of occurrences for XY pair separated by i}}{\text{Expected number of occurrences for XY pair separated by i}}$$

The LOD value is the log (base 10) of the ratio. The observed number of occurrences for any salt bridge pair is the simple count for that pair in a genome. The expected number of occurrences for that pair is calculated as follows: Given the frequencies of two amino acids X and Y in the helices as P(X) and P(Y), the probability of an XY pair occurring, assuming the occurrence of X and Y is completely uncorrelated, is:

$$P(X,Y) = P(X)\ P(Y)$$

If the total number of all amino acid pairs is N, the expected number of occurrences of the XY pair is calculated as,

$$N(XY) = N\ P(X)\ P(Y)$$

Figure 1C. Method of Stratified Resampling based on Orthologous Relationship.

This diagram illustrates our strategy of stratified resampling. A bar in the figure represents an ORF where positive-negative pairs indicate salt bridges that may be present in the protein. The first column in the figure shows various gene families. Second and third column represent the ORFs in thermophilic and mesophilic genomes. Notice that some of the ORFs are grouped into families of paralogs. The ortholog column shows the idea of stratified sampling, where we extract one representative member from each gene family for every organism. The final column indicates whether a structure of a homologous protein is available for the family. The dashed lines (-) in the figure show the sequences that are missing for any orthologous group and are thus discarded from our calculation.

More specifically, to identify orthologs we followed a five-step procedure:

(i) We started with the COGs classification at the NCBI (Tatusov, *et al.*, 1997). This currently contains 864 orthologous groups that are present in varying degrees in 8 of the first genomes sequenced (a subset of the twelve genomes used in this study).

(ii) We restricted our attention initially to the 110 COGs present in all 8 genomes.

(iii) Then we dealt with the issue of those COGs represented by multiple proteins in certain genomes (i.e. paralogs). To compensate for this effect, we chose only those COGs that had a maximum of ten sequences in total. In the few cases when we had paralogs, to pick a best representative, we consulted the dendograms on the COGs website.

(iv) To enlarge a COGs cluster to the 12 genomes used here, we performed pairwise sequence comparison using the FASTA program (version 2.0) (Lipman & Pearson, 1985) where the COGs sequences were used as queries against the four additional genomes not part of the original COGs study (i.e. AA, OT, AF, and MT). We used an ‘e-value’ threshold of 0.01 in these comparisons. The e-value describes the number of errors per query expected in a single database scan, so a value of 0.01 means that about one out of a hundred cluster linkages will be in error.

(v) Finally, we kept only those COGs that had easy-to-find members in the extra four genomes.

Application of the whole procedure resulted in the list of 52 COGs that we used in our study. A subset of 18 of these had homologs in the PDB structure databank and was used for the tertiary salt bridge study. These are indicated by the rhombus in the last column of the figure.

Figure 1D. Determination of tertiary salt bridges by an Indirect Method of Structure Mapping

To determine the positions of the salt bridges in a protein of unknown structure, where possible, we mapped its sequence onto a homologous protein of known structure in the PDB. All the salt bridges in the protein with known structure were determined by a program that takes coordinates of a protein and gives a list of hydrogen bonds as an output occurring in it (Gerstein, 1992). The list of hydrogen bonds considered here involved only side-chain/side-chain and side-chain/main-chain interactions between amino acids, as the main-chain/main-chain hydrogen bonds are mostly involved in forming secondary structural elements. Next we aligned all twelve sequences in each orthologous group with the corresponding PDB sequence by multiple sequence alignment using CLUSTALW (Higgins *et al*., 1996). Then for every salt bridge pair in the PDB protein, a corresponding amino acid pair was determined in the similar position in other proteins. It has been observed that in some proteins, the amino acid pair corresponding to a salt bridge is conserved, whereas in others it is replaced either by a non-ionic pair or by a complementary salt bridge pair.

Figure 2. Amino acid composition in genome, helix and 52 orthologous proteins.

Amino acid composition in genome-overall, helix and for orthologous proteins are shown by the additive bar graphs in 2A, 2B and 2C respectively. In the figures the blackened area represents the portion of charged residues E, D, K and R. This area increases from mesophiles to thermophiles and the trend is followed in all three levels as shown by the figures. On the contrary, the amounts of amine residues, N and Q decrease in thermophilic helices. Also note that among hydrophobic groups (AILV) there is an increase in the contents of L and V in thermophiles.

Figure 3. LOD values of the EK pair in helix as a function of separation.

LOD values for EK pair peak at a separation of either 3 or 4 suggesting that the pair at these positions represents salt bridge pair.

Figure 4A. LOD values of EK salt bridge pairs with separation of 3 and 4.

LOD values increases with the increase in physiological temperatures shown along the horizontal axis. For mesophiles, they are indicated by a range from 10 to 45 °C.

Figure 4B. Distribution of LOD values of EK(3) for randomly generated mesophilic and thermophilic genomes.

The figure shows the distribution curves of EK(3) LOD values for randomly generated thermophilic and mesophilic genomes. The difference of the two means of the distribution is $|\Delta| = 0.18$. The sample variance for thermophiles is $s_x^2 = 0.0038$ and that for mesophiles, $s_y^2 = 0.0059$. We performed a standard double-blind experiment to test the significance of the difference of means. We calculated the Z score as follows:

$$Z = \frac{<X> - <Y>}{\sigma}$$

where <X> and <Y> are the means of thermophilic and mesophilic distributions, respectively, $\sigma^2 = \mathbf{s}^2{}_x/ n_x + s^2{}_y/ n_y$, and n_x and n_y are the number of observations for each distribution (500 here). Results show that the probability that the two distributions will have same mean is less than 5 percent.

Figure 5A. Length distribution of proteins in twelve organisms.

We used an extreme value distribution for the fit curve shown by the bold line: Frequency at any protein length x is given by, y = exp(c-b(x-a)-exp(-b(x-a)) where a = 211.0, b = 0.007142, and c = 0.2277. Note that some sequences longer than 983 amino acids are not shown in the graph. Two letter abbreviations are defined in Table 1. It is evident from the figure that at shorter protein lengths thermophiles exceed the fit curve while mesophiles are below it, but at the longer protein length mesophiles exceed the fit curve and thermophiles go under.

Figure 5B. Comparison of thermophilic and mesophilic fit curves for length distribution both for overall genome sequences and orthologous proteins.

We used the same extreme value distribution for the fit curve as in Figure 6. Only the fit curves are shown here.

Figure 6A. Length distribution of proteins in terms of overall percentage composition at various length.

Amount of protein at various protein lengths shown for different genomes. In the figure vertical axis represents fraction of total amount of proteins present at various protein length for all the twelve organisms. Percentage content of proteins with longer protein length increases in yeast and decreases in archaea.

Figure 6B. Average protein length in twelve organisms.

In eukaryote category we included average protein length of protein in *C. elegans* genome (CESC, 1998). Shaded genomes represent the thermophiles. Overall averages for each category are given on the top of every category-column. Note that the average protein length for archaea is shorter than that for either of the other forms of life.

The Stability of Thermophilic Proteins: A Study based on Comprehensive Genome Comparison

Rajdeep Das

&

Mark B. Gerstein*

Department of Molecular Biophysics & Biochemistry
266 Whitney Avenue, Yale University
PO Box 208114, New Haven, CT 06520
(203) 432-6105, FAX (203) 432-5175
Mark.Gerstein@yale.edu

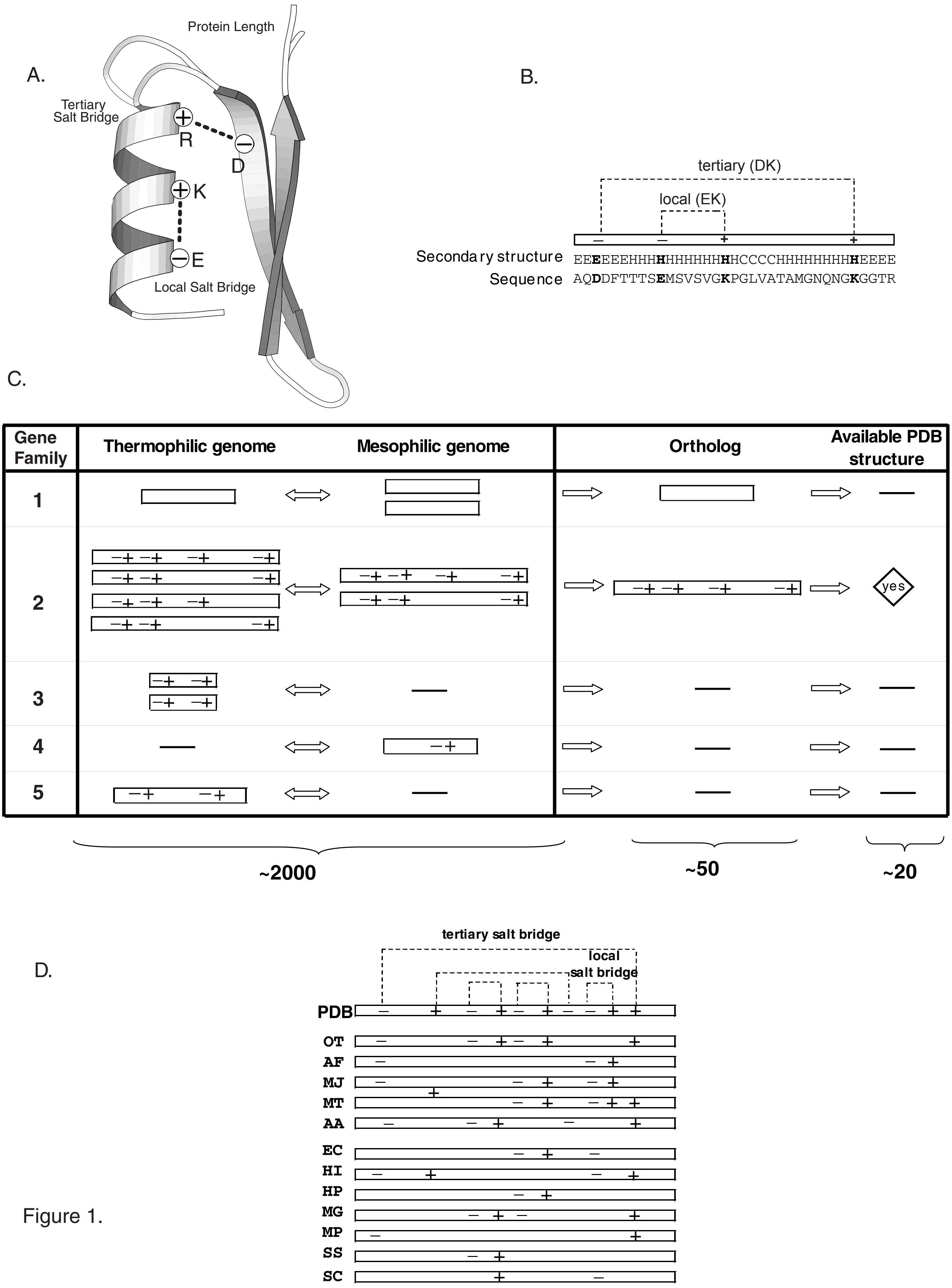
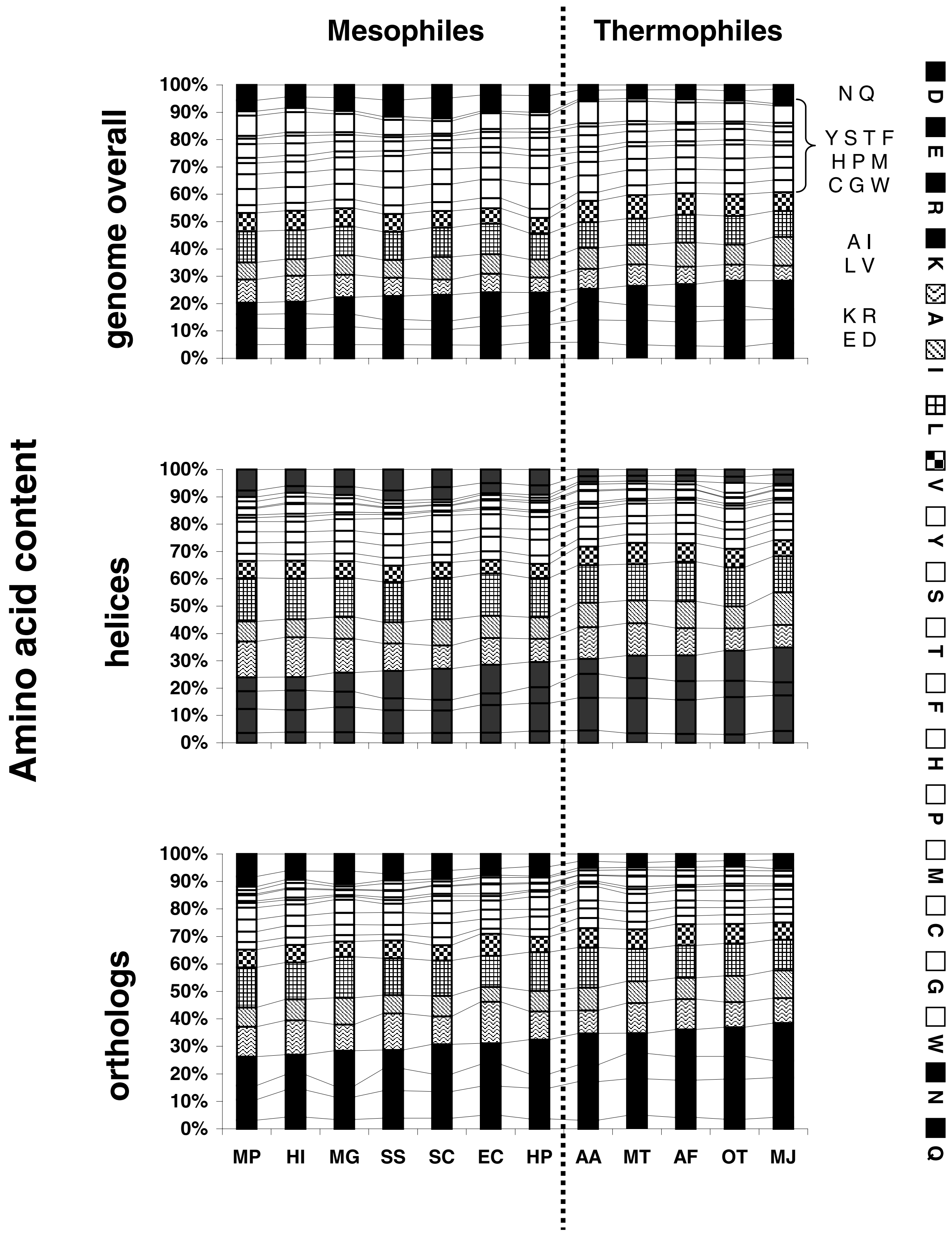

A.
Protein Length
Tertiary Salt Bridge
R
D
K
E
Local Salt Bridge
B.
tertiary (DK)
local (EK)
Secondary structure EEEEEEHHHHHHHHHHHHCCCCHHHHHHHHHEEEE
Sequence AQDDFTTTSEMSVSVGKPGLVATAMGNQNGKGGTR
C.
Gene Family
Thermophilic genome
Mesophilic genome
Ortholog
Available PDB structure
1
2
3
4
5
yes
~2000
~50
~20
D.
tertiary salt bridge
local salt bridge
PDB
OT
AF
MJ
MT
AA
EC
HI
HP
MG
MP
SS
SC

Figure 1.


Mesophiles
Thermophiles
Amino acid content
genome overall
helices
orthologs
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
MP
HI
MG
SS
SC
EC
HP
AA
MT
AF
OT
MJ
N Q
Y S T F
H P M
C G W
A I
L V
K R
E D
D
E
R
K
A
I
L
V
Y
S
T
F
H
P
M
C
G
W
N
Q

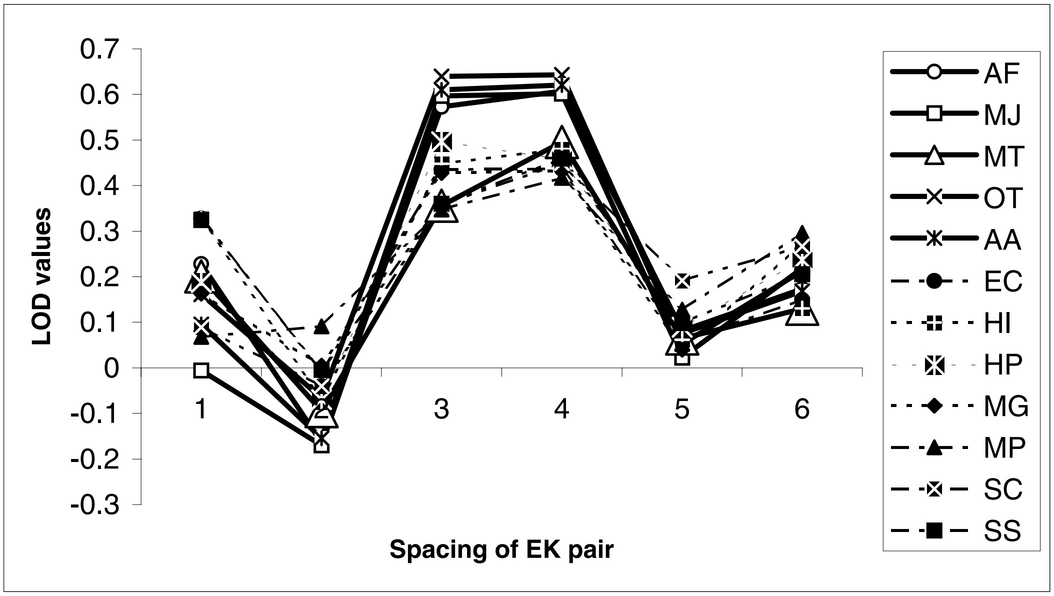

0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
LOD values
1
3
4
5
6
Spacing of EK pair
AF
MJ
MT
OT
AA
EC
HI
HP
MG
MP
SC
SS

Figure 3.

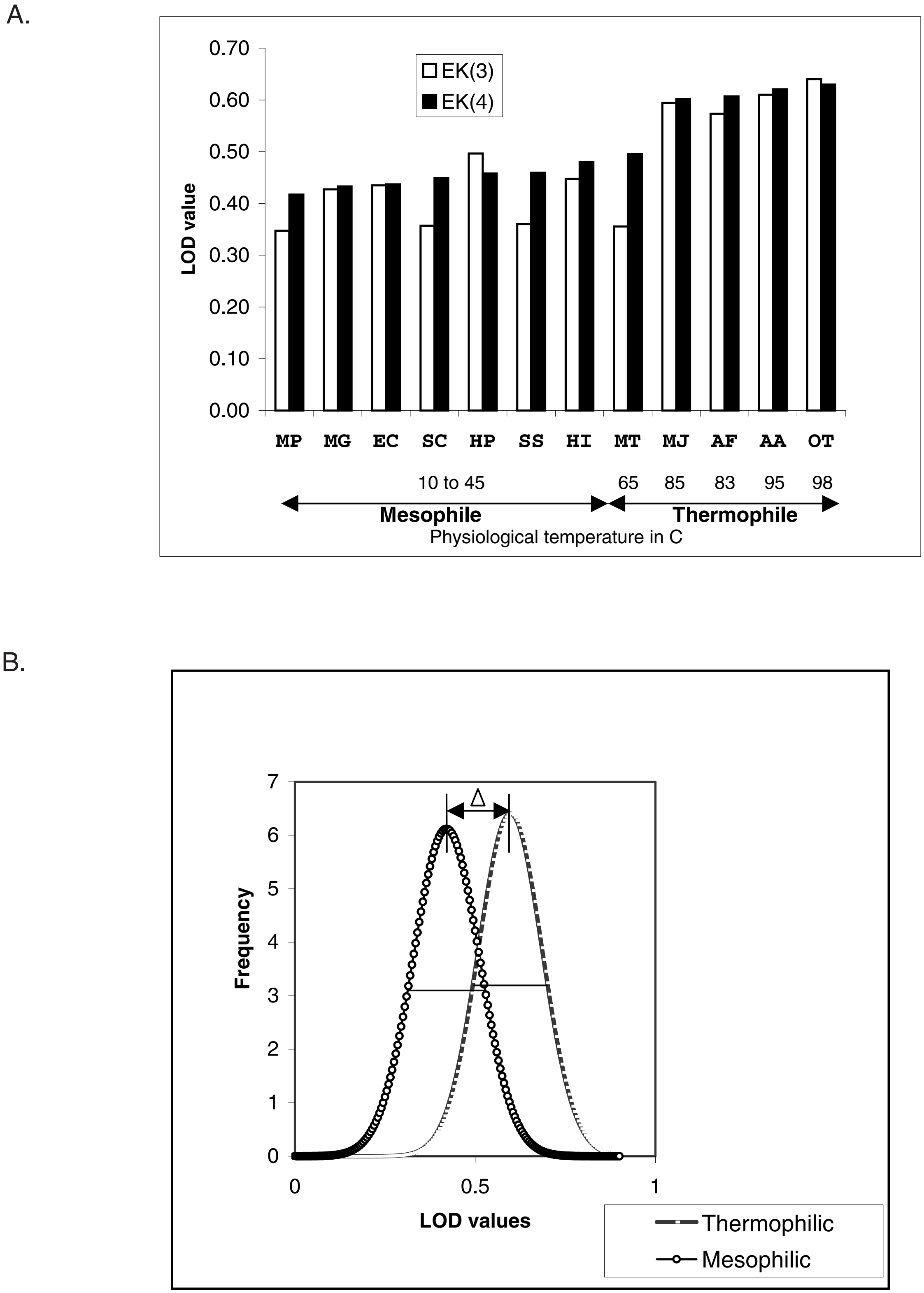
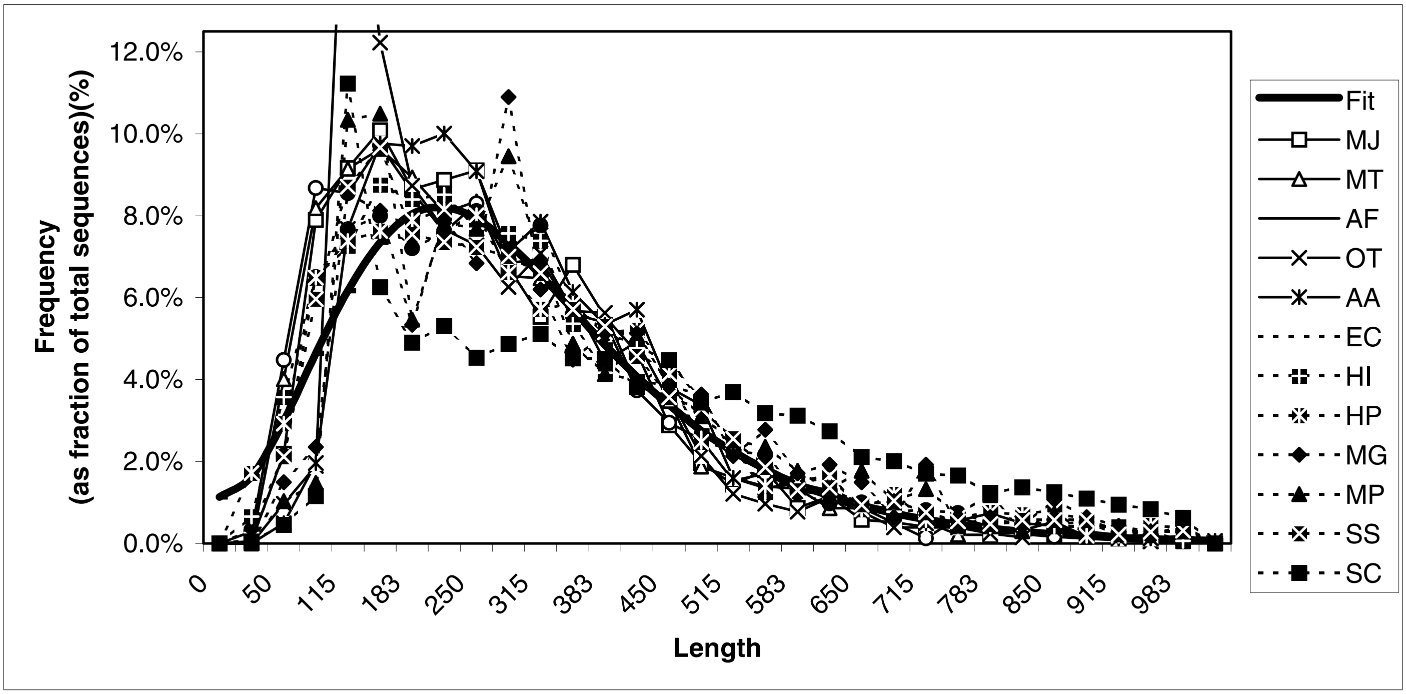

A.
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
LOD value
□ EK(3)
■ EK(4)
MP
MG
EC
SC
HP
SS
HI
MT
MJ
AF
AA
OT
10 to 45
65
85
83
95
98
Mesophile
Thermophile
Physiological temperature in C
B.
7
6
5
4
3
2
1
0
Frequency
Δ
0
0.5
1
LOD values
Thermophilic
Mesophilic

Figure 4.

A.


Frequency
(as fraction of total sequences)(%)
12.0%
10.0%
8.0%
6.0%
4.0%
2.0%
0.0%
0
50
115
183
250
315
383
450
515
583
650
715
783
850
915
983
Length
Fit
MJ
MT
AF
OT
AA
EC
HI
HP
MG
MP
SS
SC

B.

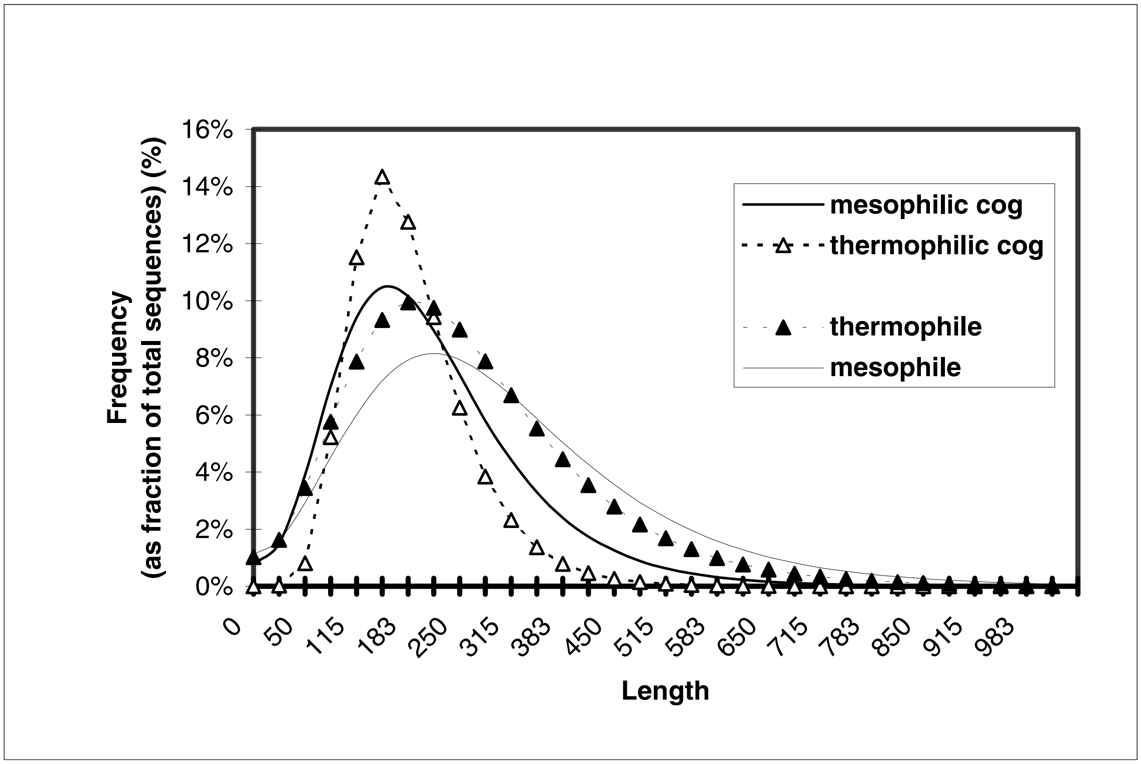
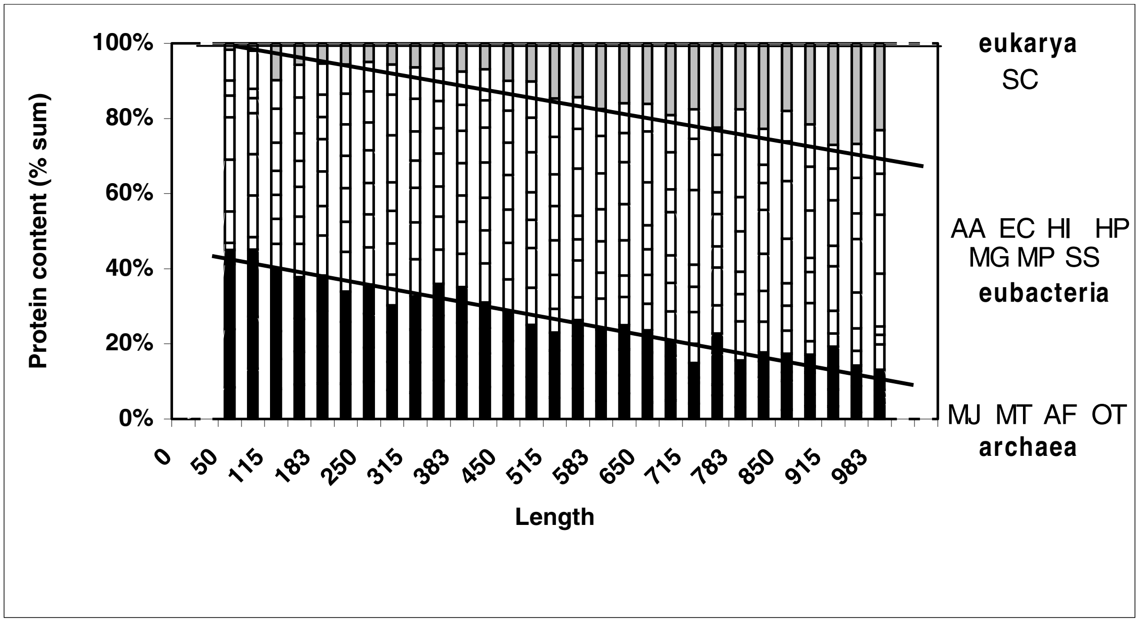

Frequency
(as fraction of total sequences) (%)
16%
14%
12%
10%
8%
6%
4%
2%
0%
0
50
115
183
250
315
383
450
515
583
650
715
783
850
915
983
Length
mesophilic cog
thermophilic cog
thermophile
mesophile

Figure 5.

A.



B.

Average	ARCHAEA ~281				BACTERIA ~326							EUKARYAOTE ~445	
Genome	MJ	MT	AF	OT	AA	EC	HI	HP	MG	MP	SS	SC	CE
Length	286	281	274	283	317	316	300	312	363	350	326	466	423

Figure 6

Organism	**Category**	**Genome ID**	**# of Proteins**	**Physiological condition**
Pyrococcus horikoshii (Strain OT3) (Kawarabayasi *et al.*, 1998)	archaea	OT	2061	98°C, anaerobe
Aquifex aeolicus (Deckert *et al.*, 1998)	eubacteria , gram negative	AA	1522	95°C
Methanococcus janaschii (Bult *et al.*, 1996)	archaea	MJ	1735	85°C, anaerobe
Archaeoglobus fulgidus (Klenk *et al.*, 1997)	archaea	AF	2409	83°C, anaerobe
Methanobacterium thermoautotrophicum (Smith *et al.*, 1997)	archaea	MT	1869	65°C, anaerobe
Haemophilus influenzae (Fleischmann *et al.*, 1995)	eubacteria, gram negative	HI	1680	mesophilic temp.
Mycoplasma genitalium (Fraser *et al.*, 1995)	eubacteria, gram positive	MG	470	mesophilic temp.
Mycoplasma pneumoniae (Himmelreich *et al.*, 1996)	eubacteria, gram positive	MP	677	mesophilic temp.
Helicobactor pylori (Tomb *et al.*, 1997)	eubacteria, gram negative	HP	1590	mesophilic temp.
Escherichia coli (Blattner *et al.*, 1997)	eubacteria, gram negative	EC	4288	mesophilic temp.
Synechocystis sp. (Kaneko *et al.*, 1996)	cyanobacteria	SS	3168	mesophilic temp.
Saccharomyces cerevisiae (Goffeau *et al.*, 1997)	eukaryote, fungus	SC	6218	mesophilic temp.

Table 1.

LOD values (%)

			Thermophile					Mesophile						
	Spacing	Pair	AF	MJ	MT	OT	AA	EC	HI	HP	MG	MP	SC	SS
A. Helix	3	EK	57	59	36	64	61	44	45	50	43	35	36	36
		ER	48	25	42	39	36	38	27	36	27	37	25	32
		DR	45	33	53	48	48	48	40	32	-3	20	30	40
	4	EK	61	60	50	63	62	44	48	46	43	42	45	46
		ER	47	27	48	46	43	36	31	23	33	25	27	33
		DR	38	38	48	44	44	44	47	48	56	60	36	42
B. Genome	3	EK	43	47	23	14	48	27	27	38	27	21	22	22
		ER	37	17	30	22	25	26	22	27	15	25	14	21
		DR	16	8	18	1	10	18	12	9	2	4	-1	11
	4	EK	40	40	28	14	40	21	24	31	19	25	23	24
		ER	31	17	31	-7	27	24	19	12	17	14	12	21
		DR	9	4	12	24	5	16	17	22	12	13	4	13
C. 52 COG Proteins	3	EK	48	57	55	38	61	38	36	45	40	23	41	26
		ER	38	-27	14	24	36	30	43	17	-5	51	-5	21
		DR	33	65	37	31	45	47	29	29	6	20	37	35
	4	EK	58	60	57	26	62	41	65	35	46	24	62	33
		ER	29	15	39	34	43	37	38	9	-2	12	38	5
		DR	54	37	76	23	32	55	52	44	60	70	52	73

Table 2.

HELIX

		Thermophile					Mesophile						
Sep	**Pair**	**AF**	**MJ**	**MT**	**OT**	**AA**	**EC**	**HI**	**HP**	**MG**	**MP**	**SS**	**SC**
3	EK	4	5	-	4	4	9	7	7	13	-	13	19
3	ER	10	-	13	18	-	12	-	14	-	-	-	-
3	DR	13	-	-	13	12	8	10	-	-	-	10	-
4	EK	5	9	12	9	7	12	9	13	11	15	10	10
4	ER	11	14	-	14	-	-	-	-	-	-	-	-
4	DR	-	-	13	13	-	9	10	11	9	7	11	18
3	DK	9	13	8	8	16	2	5	6	11	10	4	9
4	DK	10	-	16	19	-	6	11	17	-	13	15	9

GENOME

		Thermophile					Mesophile						
Sep	**Pair**	**AF**	**MJ**	**MT**	**OT**	**AA**	**EC**	**HI**	**HP**	**MG**	**MP**	**SS**	**SC**
3	EK	4	5	9	3	3	3	5	3	9	-	7	-
3	ER	6	-	4	11	11	4	7	9	-	-	9	-
4	EK	4	8	11	9	6	12	10	6	-	-	-	-
4	ER	9	-	7	14	15	3	-	-	-	-	-	-

Table 3

COG ID	**Class**	**Category**	**PDB ID**	**Therm. average #of salt bridge**	**Meso. average of salt bridge**	**Difference**	**Score**	
49	J	ribosomal	**1rss**	5.6	3.1	2.5	1	**+**
80	J	ribosomal	**1aci**	0.8	0.7	0.1	0	
81	J	ribosomal	**1ad2**	6.4	4.3	2.1	1	**+**
91	J	ribosomal	**1bxe**	1.8	0.9	0.9	0	
93	J	ribosomal	**1whi**	3	1.9	1.1	1	**+**
96	J	ribosomal	**1sei**	2	2.1	-0.1	0	
98	J	ribosomal	**1pkp**	0.6	1.7	-1.1	-1	**-**
184	J	ribosomal	**1a32**	1.8	1.9	-0.1	0	
186	J	ribosomal	**1rip**	0.4	0.9	-0.5	0	
16	J	synthetase	**1pys**	7.6	2.6	5	1	**+**
124	J	synthetase	**1ady**	9.6	6.1	3.5	1	**+**
162	J	synthetase	**2ts1**	3.8	3.3	0.5	0	
30	J	other	**1yub**	5	5.3	-0.3	0	
125	F	other	**1tmk**	0.8	0.4	0.4	0	
149	C	other	**1btm**	3	4.3	-1.3	-1	**-**
541	N	other	**1fts**	3.6	3.4	0.2	0	
112	E	other	**1cj0**	6.2	4.6	1.6	1	**+**
552	N	other	**1ffh**	4.2	4.6	-0.4	0	

Table 4.