Structural and Genomic Correlates of Hyperthermostability*

Received for publication, July 26, 2000, and in revised form, August 10, 2000. Published, JBC Papers in Press, August 11, 2000, DOI 10.1074/jbc.C000497200

Christian Cambillau‡§ and Jean-Michel Claverie‡§
From the ‡Architecture et Fonction des Macromolécules Biologiques, CNRS UPR9639, and the §Information Génétique et Structurale, CNRS UMR 1889, 31 chemin Joseph Aiguier, 13402 Marseille cedex 20, France.

While most organisms grow at temperatures ranging between 20 and 50 °C, many archaea and a few bacteria have been found capable of withstanding temperatures close to 100 °C, or beyond, such as Pyrococcus or Aquifex. Here we report the results of two independent large scale, unbiased approaches to identify global protein properties correlating with an extreme thermophilic lifestyle. First, we performed a comparative proteome analyses using 30 complete genome sequences from the three kingdoms. A large difference between the proportions of charged versus polar (noncharged) amino acids was found to be a signature of all hyperthermophilic organisms. Second, we analyzed the water accessible surfaces of 189 protein structures belonging to mesophiles or hyperthermophiles. We found that the surfaces of hyperthermophilic proteins exhibited the shift already observed at the genomic level, i.e. a proportion of solvent accessible charged residues strongly increased at the expense of polar residues. The biophysical requirements for the presence of charged residues at the protein surface, allowing protein stabilization through ion bonds, is therefore clearly imprinted and detectable in all genome sequences available to date.

The number of solved protein structures from hyperthermophiles is increasing at a fast pace (Refs. 1–13 and references therein), due in part to structural genomics approaches. Based on comparisons between mesophiles and hyperthermophiles, two main factors have been suggested as responsible for thermostability: (i) the presence of large networks of ion pairs at the protein surface (1, 8–12), (ii) protein loop shortening (7), and (iii) an increase of protein compactness (7, 12). About a dozen additional determinants have also been mentioned, but do not appear relevant to all the structures elucidated (3–6, 8, 13). Recently, it has been proposed that adaptation to high temperature might involve different mechanisms in moderate thermophiles versus extreme thermophiles (12). This follows from a markedly different temperature dependence of the hydrophobic interactions, optimal around 70 °C (14), compared with coulombic interactions, optimum at higher temperatures, up to 100 °C (9). Moderate thermophiles and extreme thermophiles should therefore be treated separately for the analysis of high temperature adaptation factors. Apparent discrepancies in the results of previous works might originate in a failure to do so.

The availability of an increasing number of complete genome sequences makes global proteome analysis a powerful tool for establishing genome-function relationships. In this respect, some differences were already noted between the amino acid compositions of subsets of ORFs1 from mesophiles and hyperthermophiles (15–18). However, previous approaches used comparisons of orthologous sequences or of structures of proteins belonging to a same family (5–8). The X-ray structures were also selected according to quality criteria. Applying these two criteria lead to a reduced number of structures analyzed and might yield to a statistically less satisfactory approach. Here, we postulated that the most basic features of protein structures, amino acid composition and localization, should be imprinted in the genome of organisms with very different lifestyles. We therefore expected that global differences should be readily detectable from a coarse but comprehensive and unbiased analysis. Here, we computed the amino acid compositions in the ORFs of all complete genome sequences available to date (30) as well as on the protein surfaces (water-accessible surface (WAS), Ref. 19) of a large set of protein structures. Although these analysis were unbiased, since no prior selection of sequences or structures was performed, they exhibit a remarkable correlation, suggesting that the replacement of polar noncharged residues by charged ones constitutes a major stabilization mechanism in the proteins of hyperthermophilic organisms.

MATERIALS AND METHODS

Data on the amino acid composition of the fully sequenced genomes were taken from the web site of Fred Tekesia, Pasteur Institute (Paris). Organisms have been classified according to the three major domains of life and using the following abbreviations (which are used in Figs. 1 and 2A). Mesophiles (total of 22): (i) Eukaryotes, Saccharomyces cerevisiae (SC), C. elegans (CE), Schizosaccharomyces pombe (SP; 68% of total proteome), Drosophila melanogaster (DM), Arabidopsis thaliana (ATH; not complete), Plasmodium falciparum (PF; not complete); (ii) Bacteria, Mycoplasma genitalium (MG), Mycoplasma pneumoniae (MP), Synecochystis sp. (Spp.), hemeophilia influenzae (HI), Escherichia coli (EC), Helicobacter pylori (HP), Bacillus subtilis (BS), Borrelia burgdorfi (BB), Mycobacterium tuberculosis (MT), Treponema pallidum (TP), Chlamydia trachomatis (CT), Chlamydia pneumoniae (CP), Rickettsia prowazekii (RP), Campylobacter jejuni (CJ), Deinococcus radiodurans (DR), Neisseria meningitidis (NM). Thermophile (total of 1): Archaea, Methanobacterium thermoautotrophicum (MTH). Hyperthermophiles (total of 7): (i) Bacteria, Aquifex aeolicus (AE), Thermotoga maritima (TM); (ii) Archaea, Pyrococcus furiosus (PF), Pyrococcus abyssi (PA), Methanococcus jannaschii (MJ), Archaeoglobus fulgidus (AF), Aeropyrum pernix K1 (APE).

The three-dimensional structures come from the Protein Data Bank at RCSB. Water-accessible surface calculations were performed with the program DSSP (22). The number of Protein Data Bank structures used for surface calculations are as follows. (i) Mesophiles: E. coli, 50; B. subtilis, 34; S. cerevisiae, 47; (ii) Hyperthermophiles (total of 58): Pyrococcus furiosus, 11; woei, 3; kodakarenensis, 2; T. maritima, 10; neutolitica, 1; Aquifex pyrophilus, 2; eolicus, 1; Sulfolobus solfataricus, 6; acidocaldarius, 5; M. jannaschii, 9; Thermococcus litoralis, 2; eoler, 1; gordonarius, 1; Methanopyrus kandleri, 2; Pyrobaculum aerophilum, 2.

---

* The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. Section 1754 solely to indicate this fact.

§ To whom correspondence should be addressed. Fax: 33-491-16-45-36; E-mail: cambillau@afmb.cnrs-mrs.fr.

1 The abbreviations used are: ORF, open reading frame; WAS, water-accessible surface; Ch-Po, the difference between charged and polar noncharged amino acids.
Table I

Optimal growth temperature of the thermophile and the different hyperthermophiles used in this work

| Hyperthermophile                  | Bacteria | Thermophile                      | Archaea | Thermophile                     |
|-----------------------------------|----------|----------------------------------|---------|----------------------------------|
| Aquifex pyrophilus                | 85 °C    | Aquifex neobius                   | 90 °C   | Methanothermobacter thermautotrophicus | 65 °C |
| Thermotoga maritima               | 80 °C    | Archeoglobus fulgidus            | 85 °C   | Pyrobaculum aerophilum           | 98 °C |
| Aeropyrum pernix                   | 90 °C    | Methanococcus jannaschii         | 85 °C   | Pyrococcus abyssii               | 97 °C |
| Methanopyrus kandleri             |          | Methanopyrus maripaludis        |         | Pyrococcus furiosus              | 97 °C |
| Pyrococcus hortitibii             |          | Sulfolobus solfataricus          | 85 °C   | Thermococcus litoralis           | 83 °C |
| Thermophile                        |          | Thermobacter thermautotrophicus  | 65 °C   | Thermophile                      |       |

RESULTS AND DISCUSSION

Amino Acid Composition of the Proteomes—Our data set consists of the amino acid compositions of the entire proteomes of 22 mesophiles, 1 thermophile, and 7 hyperthermophiles. We calculated the average proportion of each amino acid in all mesophiles (30–50 °C) on one hand and in all hyperthermophiles (>80 °C, Table I) on the other hand. The only moderately thermophile organism genome available, that of M. thermautotrophicum (Table I), was excluded from the comparisons for the reasons mentioned above (except for Figs 1 and 2A). We then grouped the results for each amino acid in classes such as charged, polar noncharged, aliphatics, and aromatics, a procedure based on biophysical characters and already used in this kind of approach (5, 12). A benefit of this procedure is also to correct the bias in the proportion of certain amino acids induced by the strong variation in (G + C) content across the compared genomes (Fig. 1). For instance, arginine content was found to be strongly correlated with the (G + C) percentage, while the lysine content is strongly anticorrelated (Fig. 1). The use of Lys or Arg by a given organism does not depend on specific chemical properties of Arg versus Lys, but only depends on the (G + C) content, as indicated by Fig. 1. Therefore, grouping Lys + Arg makes it possible to overcome the (G + C) content bias. As a consequence, the variations observed for the positively charged residues (Lys + Arg) stay within 50%, while up to 5-fold variations are observed for the individual amino acids (Fig. 1). It is worth to notice that the (G + C) content is not correlated to the optimal growth temperature of the organisms (Fig. 1).

The proteome analysis in amino acid classes indicated that hyperthermophilicity is characterized by a sharp increase of charged residues, Lys and Glu, at the expense of polar noncharged residues, mainly Gln (Fig. 2A). This correlation with growth temperature is verified across the three kingdoms (Fig. 2A). A slight increase in medium and large aliphatics is also observed, while the percentage of the short aliphatic alanine is lower (Fig. 2B). No significant variations are observed in aromatics or in other residues such as His, Pro, Gly, or Cys. In summary, the difference between charged and polar noncharged amino acids (Ch-Po) is the best indicator of the organism’s lifestyle (Fig. 2, A–C). A threshold is observed around 10% for the extreme thermophiles Ch-Po values, while a higher limit of 5% characterizes the mesophiles. The moderate thermophile M. thermautotrophicum is situated in between these two values (Fig. 2A).

The Case of A. pernix and P. horikoshii—Our initial analysis of the A. pernix proteome revealed a discrepancy: a Ch-Po value of only 5.2% was computed for this organism, thus in the mesophile range and much below the 10% lower limit inferred from the other true hyperthermophiles (Table I). This conflicting result prompted us to reassess the quality of the ORF annotation originally proposed for this genome. According to the original publication (20) A. pernix was found to exhibit 2694 ORFs for a total genome length of 1,669,695 base pairs, thus corresponding to 1 ORF for 620 nucleotides in average. This theoretical gene density is almost twice as much as previously reported for any other micro-organism sequenced so far. The fact that this organism was also claimed to exhibit the largest proportion of ORFs without data base similarity (57.1%) suggested an anomaly in the genome annotation. The whole A. pernix ORF data set was then downloaded from the official site at the Japan National Institute for Technology Evaluation for further study. It soon appeared that the total number of nucleotides included in the protein coding fraction of the genome (1,916,052) was much greater than the total number of nucleotides (1,669,695). This anomaly is mostly due to the annotation of overlapping ORFs in different reading frames, resulting in a significant overprediction of genes and the pollution of our computations by illegitimate protein sequences.

We first corrected this problem by limiting our amino acid composition analyses to the subset of 628 predicted ORFs for which clear homologues have been recognized in other organisms (“Annotable ORFs,” according to Ref. 20). This subset is less likely to contain illegitimate ORFs. Independently, we also reannotated A. pernix genomic sequence using the iterative Markov model method SelfID (21) and recomputed the amino acid composition on the predicted coding regions longer than 300 nucleotides. The amino acid compositions obtained by these two independent ways are virtually identical, and expected, very different from the composition computed on the original data set. Using the two corrected amino acid compositions, our previously defined indicator of thermophilicity Ch-Po was 10.6%, compatible with the values observed for the other hyperthermophilic organisms.

After the above correction for A. pernix, P. horikoshii became the hyperthermophile organism associated with the lowest
Ch-Po value (10.2%), far from the other Pyrococcus, *P. abyssii* (13.2%). It turns out that the *P. horkoshii* genome sequence was determined in the same Institute as that of *A. pernix* and *FIG. 3. Amino acid water-accessible surface area within the 189 protein structures under study. A, percentages of the various amino acids WAS (relative to the total surface area calculated for each Protein Data Bank file) belonging to mesophiles (blue) and hyperthermophiles (red) plotted versus the 20 amino acids. B, percentages of the WAS of the different classes of amino acids for three mesophiles (yeast, *B. subtilis*, and *E. coli*) and their average (blue) compared with three hyperthermophiles (*Thermococcus*, *Pyrococcus*, and *Methanococcus*) and their average (red) (see "Materials and Methods").

Ch-Po value (10.2%), far from the other Pyrococcus, *P. abyssii* (13.2%). It turns out that the *P. horkoshii* genome sequence was determined in the same Institute as that of *A. pernix* and

**Fig. 2.** Amino acid content variations within the 30 genomes under study (see Fig. 1). A, plot of the sum of the percentages of charged amino acids (Lys, Arg, Asp, Glu; *CHA*, blue), polar noncharged amino acids (Asn, Gln, Ser, Thr; *POL*, green), and of the difference of the two values (*CH-POL*, red). The mesophiles and hyperthermophiles are identified by MESO and HYPER, respectively. The eukaryotes are identified. All the other mesophiles are bacteria, while most hyperthermophiles are archaea except the two bacteria (ae and tm) identified by black circles. The sole moderate thermophile is the archaea *M. thermoautotrophicum* (*mth*, isolated red square). B, plot of the percentages of the various amino acids in mesophiles (blue) and hyperthermophiles (red), also identified in their class. C, plot of the sum of percentages of the various amino acid classes in mesophiles (blue) and hyperthermophiles (red).
annotated according to a similar protocol (23). *P. horkoshii* is also characterized by a very large fraction of ORFs without data base similarity (44.4%) and a unusually high gene density with predicted protein coding nucleotides representing 98.5% of the total genome. We thus suspected that the same overprediction plaguing *A. pernix* could affect *P. horkoshii*, although to a lesser extent. As expected, restraining the amino acid composition analysis to the subset of 557 predicted ORFs with recognized homologs resulted in a corrected C-P value of 13.6%, almost identical to that of *P. abyssi*. The average values presented in the plots of Fig. 2 include these corrections.

In the paper by Audic and Claverie (21) 10 bacterial genomes were checked using the described method. There were very small differences between the published data and the reanalyzed one. After the discovery of the wrong annotations of the Japanese institute, we reanalyzed the 13 remaining genomes, and we found, as in the first 10, small differences or none.

The Water-accessible Surfaces of Amino Acid Classes—The proportion of polar noncharged amino acids is greatly decreased in the extreme thermophile organisms and partially compensated by charged amino acids (Fig. 2A). Where are these amino acids located in the proteins structures? Are the variations observed in the genomes correlated with differences at the protein surfaces? To address these questions, we calculated the WAS of the various amino acids for 131 proteins from three mesophilic bacteria and for 58 proteins from hyperthermophilic bacteria or archaea (Fig. 3). The WAS values were then grouped by classes of amino acids, as defined above for the genomic analyses. The most striking difference between the WAS values was again found between charged and polar noncharged amino acids (Fig. 3A). The WAS percentage of charged amino acids increased from 22.5 in mesophiles to 27 in hyperthermophiles (+20%), mainly due to lysines and glutamic acids. This difference (+4.5%) was almost fully compensated for by the decrease of the WAS percentage (−3.5%) of polar residues in mesophiles (Fig. 3B). Glutamine exhibited the largest variation among all residues, with a WAS decreased by more than 2-fold. WAS values for the other polar residues exhibited the same trend, to a lesser magnitude. The WAS percentage for aliphatic residues showed little change at 6.0–6.5% of the total surface area. In contrast, alamines are decreased by 2-fold, and histidines by 1.5-fold (Fig. 3A). Finally, the exposed area of Gly, Cys, and aromatic residues was found to be unchanged between mesophiles and hyperthermophiles. The differences observed in the above average values are comparable with those computed between individual sets of proteins from the three mesophiles or the three hyperthermophilic organisms, reinforcing their significance (Fig. 3B). Thus, our results indicate that the large differences in surface accessibilities of amino acid classes are correlated with the large variations observed in the overall proteome abundance between the two sets of organisms.

**Concluding Remarks**—Coarse but extensive approaches at the genome or structure level are thus able to identify discriminant factors between mesophiles and extreme thermophiles. This implies the existence of global structural features associated with hyperthermostability common to thermophilic bacteria and archaea. The equal increase of subclasses of oppositely charged residues (mostly Lys, Arg, and Glu) in hyperthermophiles most likely derives from the increased amount of ion pairs observed at the surface of their proteins. This is the genetic signature of a thermodynamic advantage resulting from the increased stability of coulombic interactions with temperature (itself due to the decrease of water dielectric constant, Ref. 9). Furthermore, in most structures of hyperthermophilic proteins, the existence of long chains of ion pairs providing cooperative stabilization has been demonstrated (1–3).

Interestingly, the residues the proportion of which exhibit the largest relative decreases in the genomes of hyperthermophiles are Asn and Gln, already known to be the most temperature sensitive. It is tempting to relate the observed Glu/Gln balance to a peculiarity in the metabolism of hyperthermophiles. For instance, the Gln tRNA synthetase gene is missing in the bacteria *A. aeolicus* (18) and the archaea *M. janaschii* (24). In both organisms, Gln originates from the amidation of glutamic acid loaded on the Gln tRNA. Conversely, a completely different class of Lys tRNA synthetase gene has been found to exist in some hyperthermophiles, where the proportion of lysine is much increased. Missing, different, or more efficient aminoacyl tRNA synthetases may contribute to various adaptive mechanisms and play a crucial and general role in the pathway to hyperthermostability.

**Acknowledgments**—The computing assistance of Stephane Audic and Eric Blanc is greatly acknowledged.

**REFERENCES**

1. Goldman, A. (1995) *Structure (Lond.)* 3, 1277–1279
2. Korolev, S., Naylor, M., Barnes, W. M., Di Cera, E., and Waksman, G. (1995) *Proc. Natl. Acad. Sci. U. S. A.* 92, 9264–9268
3. Yip, K. S., Stillman, T. J., Britton, K. L., Artymiuk, P. J., Baker, P. J., Sedelnikova, S. E., Engel, P. C., Pasquoto, A., Chiarulli, R., and Consalvi, V. (1995) *Structure (Lond.)* 3, 1147–1158
4. Russel, R. J. M., Ferguson, J. M. C., Hough, D. W., Danson, M. J., and Taylor, G. L. (1997) *Biochemistry* 36, 9883–9994
5. Vogt, G., Wuehl, S., and Argos, P. (1997) *J. Mol. Biol.* 269, 631–643
6. Lo Leggio, L., Kalogiannis, S., Bhat, M. K., and Pickersgill, R. W. (1999) *Proteins Struct. Funct. Genet.* 36, 285–296
7. Thompson, M. J., and Eisenberg, D. (1999) *J. Mol. Biol.* 290, 595–604
8. Mass, D., Zeelen, J. P., Thanki, N., Beaucamp, N., Alvarez, M., Thi, M. H., Backmann, J., Martial, J. A., Wierenga, R. K. (1999) *Proteins Struct. Funct. Genet.* 37, 441–453
9. Pace, N. (2000) *Nat. Struct. Biol.* 7, 345–346
10. Perl, D., Mueller, U., Heineman, U., and Schmid, P. X. (2000) *Nature Struct. Biol.* 7, 389–381
11. Pace, N., Alston, R., and Shaw, K. L. (2000) *Protein Sci.*, in press
12. Szilagyi, A., and Zavodszky, P. (2000) *Structure (Lond.)* 8, 495–504
13. Danzon, M. J., and Hough, D. W. (1998) *Trends Microbiol.* 6, 307–314
14. Scheraga, H. A. (1998) *J. Biomol. Struct. Dyn.* 16, 447–460
15. Tekaia, F., Lazcano, A., and Dujon, B. (1999) *Genome Res.* 9, 550–557
16. Bohm, G., and Jaenicke, R. (1994) *Int. J. Pept. Protein Res.* 43, 97–106
17. Choi, I.-G., Kim, S. S., Ryu, J. R., Han, Y. S., Bang, W. G., Kim, S. H., and Yu, Y. G. (1997) *Extremophiles* 1, 125–134
18. Deckert, G., et al. (1998) *Nature* 392, 353–358
19. Connolly, M. L. (1983) *Science* 219, 287–290
20. Kawarabayasi, Y., et al. (1999) *DNA Res.* 6, 83–101
21. Audic, S., and Claverie, J.-M. (1998) *Proc. Natl. Acad. Sci. U. S. A.* 95, 10266–10303
22. Kahveci, W., and Sander, C. (1985) *Biopolymers* 22, 2577–2637
23. Kawarabayasi, Y., et al. (1998) *DNA Res.* 5, 55–76
24. Bult, C. J., et al. (1997) *Science* 273, 1058–1073