J. Biochem. **133**, 507–513 (2003)
DOI: 10.1093/jb/mvg067

Compositional Changes in RNA, DNA and Proteins for Bacterial Adaptation to Higher and Lower Temperatures

Hiroshi Nakashima*[,1], Satoshi Fukuchi[2] and Ken Nishikawa[2]

[1]School of Health Sciences, Faculty of Medicine, Kanazawa University, 5-11-80 Kodatsuno, Kanazawa 920-0942; and [2]Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Yata 1111, Mishima, Shizuoka 411-8540

Received November 21, 2002; accepted January 29, 2003

It is known that in thermophiles the G+C content of ribosomal RNA linearly correlates with growth temperature, while that of genomic DNA does not. Although the G+C contents (singlet) of the genomic DNAs of thermophiles and methophiles do not differ significantly, the dinucleotide (doublet) compositions of the two bacterial groups clearly do. The average amino acid compositions of proteins of the two groups are also distinct. Based on these facts, we here analyzed the DNA and protein compositions of various bacteria in terms of the optimal growth temperature (OGT). Regression analyses of the sequence data for thermophilic, mesophilic and psychrophilic bacteria revealed good linear relationships between OGT and the dinucleotide compositions of DNA, and between OGT and the amino acid compositions of proteins. Together with the above-mentioned linear relationship between ribosomal RNA and OGT, the DNA and protein compositions can be regarded as thermostability measures for RNA, DNA and proteins, covering a wide range of temperatures. Both the DNA and proteins of psychrophiles apparently exhibit characteristics diametrically opposite to those of thermophiles. The physicochemical parameters of dinucleotides suggested that supercoiling of DNA is relevant to its thermostability. Protein stability in thermophiles is realized primarily through global changes that increase charged residues (*i.e.*, Glu, Arg, and Lys) on the molecular surface of all proteins. This kind of global change is attainable through a change in the amino acid composition coupled with alterations in the DNA base composition. The general strategies of thermophiles and psychrophiles for adaptation to higher and lower temperatures, respectively, that are suggested by the present study are discussed.

Key words: amino acid composition, DNA dinucleotides, optimal growth temperature, psychrophiles, regression analysis, thermophiles.

Abbreviations: OGT, optimal growth temperature, BLAST, basic local alignment search tool; PSI-BLAST, position-specific iterated BLAST; Pfam, protein families database of alignments and hidden Markov models.

Microorganisms that grow above 55 degrees Celsius (°C) and below 20°C are called thermophiles and psychrophiles, respectively, the remainder being called mesophiles. Extreme thermophiles can tolerate 100°C and extreme psychrophiles can survive at nearly 0°C. All macromolecules of such bacteria, *e.g.*, RNA, DNA and proteins, must be stable and functional in the temperature range in which the species lives. Much research has been carried out to elucidate the mechanisms of bacterial adaptation to higher and lower temperatures. As G·C pairs held together with three hydrogen bonds are more stable than A·T pairs with two bonds, the G+C contents of thermophilic genomes are expected to be higher. In fact, the G+C content of ribosomal 16S RNA has been reported to be proportional to the bacterial growth temperature (*1*, *2*). However, no such simple relationship exists for genomic DNA (*3*); the average G+C content varies in the wide range from approximately 30% to more than 60% in various bacteria irrespective of the thermophilic/mesophilic types (see Table 1). It has been reported that the synonymous codon usage in genes of thermophiles is different from that of mesophiles (*4*). A recent report by Kawashima *et al*. (*5*) demonstrated that not the G+C content, but the combination of purine/pyrimidine dinucleotide compositions correlates linearly with the optimal growth temperature (OGT) among archaea. A number of comparative studies on the three-dimensional structures of proteins between thermophilic and mesophilic homologues have revealed some molecular determinants for adaptation to higher temperatures. The factors hitherto suggested to be responsible for thermal stability are increased occurrences of electrostatic interactions (*6*), hydrogen bonds (*7*), and deletions in exposed loop regions (*8*). Reportedly, thermophilic and mesophilic proteins are distinctive in their amino acid compositions (*8*, *9*), the difference being mainly due to dissimilar surface compositions (*10*, *11*).

Motivated by the above mentioned study by Kawashima *et al*. (*5*), we have conducted regression analyses of DNA

*To whom correspondence should be addressed. Tel.: +81-76-265-2582, Fax: +81-76-234-4360, E-mail: naka@kenroku.kanazawa-u.ac.jp

Table 1. **List of species used in this work.**

Species	G+C (%)	Genome size (Mbp)	OGT (°C)	Reference
Thermophilic archaea				
Methanococcus jannaschii	31.4	1.7	85	Bult *et al.* (*22*)
Salfolobus tokodaii	32.8	2.7	80	Kawarabayasi *et al.* (*23*)
Salfolobus solfataricus	35.8	3.0	80	She *et al.* (*24*)
Thermoplasma volcanium	39.9	1.6	60	Kawashima *et al.* (*25*)
Pyrococcus horikoshii	41.9	1.7	98	Kawarabayasi *et al.* (*26*)
Pyrococcus abyssi	44.7	1.8	103	unpublished
Thermoplasma acidophilum	46.0	1.6	59	Ruepp *et al.* (*27*)
Archaeoglobus fulgidus	48.6	2.2	83	Klenk *et al.* (*28*)
Methanobacterium thermoautotrophicum	49.5	1.8	65	Smith *et al.* (*29*)
Aeropyrum pernix	56.3	1.7	95	Kawarabayasi *et al.* (*30*)
Thermophilic eubacteria				
Aquifex aeolicus	43.5	1.6	85	Deckert *et al.* (*31*)
Thermotoga maritima	46.2	1.9	80	Nelson *et al.* (*32*)
Mesophilic eubacteria				
Campylobacter jejuni	30.5	1.6	43	Parkhill *et al.* (*33*)
Mycoplasma genitalium	31.7	0.6	37	Fraser *et al.* (*34*)
Haemophilus influenzae	38.1	1.8	37	Fleishmann *et al.* (*35*)
Helicobacter pylori	38.9	1.7	37	Tomb *et al.* (*36*)
Mycoplasma pneumoniae	40.0	0.8	37	Himmelreich *et al.* (*37*)[a]
Bacillus subtilis	43.5	4.2	37	Kunst *et al.* (*38*)
Yersinia pestis	47.6	4.7	37[a]	Parkhill *et al.* (*39*)
Escherichia coli	50.8	4.6	37	Blattner *et al.* (*40*)
Mycobacterium tuberculosis	65.6	4.4	37	Cole *et al.* (*41*)
Pseudomonas aeruginosa	66.6	6.3	37	Stover *et al.* (*42*)
Psychrophiles				
Cenarchaeum symbiosum etc.			10	Preston *et al.* (*43*)

[a]Cultivation temperature

dinucleotide compositions as well as protein amino acid compositions against bacterial OGT, using data for thermophiles, mesophiles and psychrophiles, encompassing both archaea and eubacteria. We also reexamined the relationship between the G+C content of RNA and OGT.

MATERIALS AND METHODS

Table 1 lists 12 thermophilic and 10 mesophilic bacteria whose complete genomes have been determined and whose OGTs are known (http://www.dsmz.de/species/strains.htm). As there are no psychrophilic bacteria whose genomes have been entirely sequenced, their sequence data were treated differently (see below). A large amount of data is needed to produce reliable results on regression analysis. Therefore, we selected 100 genes or proteins from each species of the wholly sequenced mesophilic and thermophilic bacteria in the GTOP database (*12*) (http://spock.genes.nig.ac.jp/~genome/gtop.html). This database provides not only 3D structural-assignments with a PSI-BLAST search against the Protein Data Bank, but also other standard search results, such as with BLAST, Pfam, ProSite, SOSUI (*13*) (for transmembrane domain prediction), and SignalP (*14*) (for prediction of secretion signal peptides), for all proteins encoded by the genomes it contains.

The selection criteria used were as follows. First, we excluded all probable membrane proteins detected by SOSUI as well as probable extracellular proteins identified by SignalP in order to choose only typical intracellular proteins. All the remaining proteins of each bacterium were sorted by length and then divided into 100 sections, and one protein was picked from each section. The nucleotide sequences corresponding to the proteins were stored together with the sample data. Twenty-three protein sequences of psychrophilic species were taken from the SwissProt database (*15*), and the corresponding nucleotide sequences were obtained from the DDBJ/EMBL/GenBank database (*16*). In total 2,223 protein sequences and corresponding nucleotide sequences were selected for this work. Sequence data for ribosomal RNAs (rRNA) and transfer RNAs (tRNAs) were obtained according to the annotations of genome data, although annotations for rRNAs and tRNAs of *P. abyssi* were unavailable.

We performed least-square regression analysis on the DNA data to find relationships between bacterial OGT and ten symmetrical components (*i.e.*, AT, TA, GC, CG, AA/TT, AG/CT, GA/TC, GG/CC, AC/GT and CA/TG) of dinucleotide compositions. Similarly, the relationships between amino acid compositions and OGT were investigated using the protein data.

RESULTS

G+C Contents of 16S rRNAs Versus Temperature—The G+C contents of 5S, 16S and 23S rRNAs, and tRNA of all bacteria examined were found to exhibit strong correlations with the optimal growth temperature (OGT), in agreement with previous reports (*1*, *2*). The G+C content of 16S rRNA (filled circles in the top panel of Fig. 1) of thermophiles exhibited the best linear correlation with

OGT (broken line, correlation coefficient: 0.97). For thermophiles, the relationship between G+C content and OGT can be expressed by the equation,

$$\text{OGT-RNA} = 2.91 \times (\text{G+C}) - 103 \quad (1)$$

where OGT-RNA is OGT estimated in degrees Celsius (°C), and G+C refers to the percentage guanine and cytosine occupy in 16S rRNA. The G+C contents of 16S rRNAs from mesophiles and psychrophiles, on the other hand, show greater deviations from the line, ranging from 45 to 60%. Thus, the entire plot of closed circles appears biphasic, *i.e*., a linear OGT dependence of thermophiles, and OGT-independent variation of mesophiles and psychrophiles. In fact, the variations in mesophiles depend on the G+C content of genomic DNA (*17*), whereas no correlation between the G+C contents of RNA and genomic DNA was observed in thermophiles (see below).

In the same panel, the G+C contents of tRNA (open triangles) and protein coding gene (open circles) are also plotted for thermophiles. It is interesting to note that the G+C contents of tRNAs are always a little higher than those of 16S rRNAs. There was no difference between the two groups of tRNAs, one belonging to the ribosomal operons and the other not functioning as ribosomal operons. In contrast to 16S rRNAs and tRNAs, the genomic DNA of thermophiles showed no correlation with OGT, indicating that the G+C content *per se* does not contribute to the thermal stability of genomic DNA.

Dinucleotide Composition of DNA—Kawashima *et al*. (*5*) reported that in archaea a simple combination of purine (R) and pyrimidine (Y) dinucleotide compositions, RR + YY – RY – YR, is linearly correlated with OGT. In our data set, the same variables gave a correlation coefficient of 0.66. The combination of Eq. 2 obtained by least-square regression analysis exhibited better linearity with a correlation coefficient of 0.86 between the calculated and real OGTs for the 2,223 sequences. Calculated optimal temperature (OGT-DNA) is expressed in terms of the ten symmetrical components of the dinucleotide composition as,

$$\begin{aligned}\text{OGT-DNA} = {} & 0.22(\text{AA/TT}) + 4.64(\text{AG/CT}) + 1.17(\text{GA/TC}) \\ & + 3.53(\text{GG/CC}) - 2.86(\text{AC/GT}) - 0.63(\text{CA/TG}) \\ & - 0.55(\text{AT}) + 2.30(\text{TA}) - 7.09(\text{GC}) + 2.52(\text{CG}) - 0.89 \quad (2)\end{aligned}$$

The middle panel of Fig. 1 presents a plot of OGT-DNA using the average composition of each bacterial species versus real OGT (correlation coefficient: 0.93). The relationship holds in a wide temperature range covering not only thermophiles but also mesophiles, whereas the average point for psychrophiles deviates upward from the line.

In order to see how effectively the estimation with Eq. 2 holds for individual genes, we plotted histograms of the OGT-DNA values for the 100 representative genes of both *E. coli* (mesophile) and *A. pernix* (thermophile) in Fig. 2. Clear separation of the two distributions indicates that Eq. 2 is applicable to individual genes. It is remarkable that genes in thermophilic bacteria are uniformly stabilized. Therefore, it is not the mononucleotide composition such as the G+C content, but the dinucleotide composition that is indicative of the thermostability of DNA.

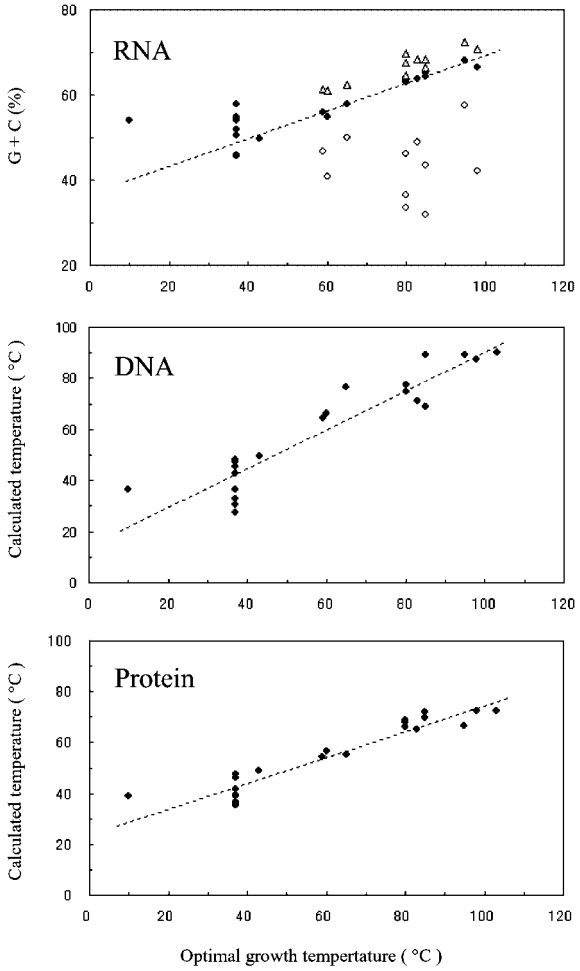

Fig. 1. **Linear correlations of RNA, DNA, and proteins with bacterial OGT.** The G+C contents of 16S rRNAs (filled circles) and tRNAs (open triangles) are plotted against OGT (labeled "RNA," top panel). Open circles represent the average G+C contents of the protein coding genes of the thermophilic species examined. The values of OGT-DNA calculated with Eq. 2 in the text are plotted against the real OGT of individual bacteria ("DNA," middle panel). The regression line (dotted line) was actually obtained by using the dinucleotide composition data for 2,223 genes, and each spot in the figure represents the average of the calculated temperatures of all genes within a single bacterium. Genes of psychrophiles are indicated by a single point representing the average, although they come from several different species. OGT-AA given by Equation (3) in the text is plotted against the real OGT of individual bacteria ("Protein," bottom panel). The regression line (dotted line) was obtained by using the amino acid composition data for 2,223 proteins, and each point in the figure represents the average of the calculated temperatures of all proteins within a single bacterium. The single point for the average value for psychrophiles represents proteins of several different species.

Amino Acid Composition of Proteins—The relationship of amino acid compositions with OGT we found in least-square regression analysis is

$$\begin{aligned}\text{OGT-AA} = {} & -0.96\text{Ala} - 0.85\text{Cys} - 2.57\text{Asp} + 1.77\text{Glu} + \\ & 0.64\text{Phe} + 0.63\text{Gly} - 1.79\text{His} + 2.60\text{Ile} + 1.22\text{Lys} + \\ & 1.26\text{Leu} + 0.62\text{Met} - 1.27\text{Asn} + 1.49\text{Pro} - 3.51\text{Gln} + \\ & 1.37\text{Arg} - 0.83\text{Ser} - 0.48\text{Thr} + 2.10\text{Val} + 1.95\text{Trp} + \\ & 2.53\text{Tyr} + 0.45 \quad (3)\end{aligned}$$

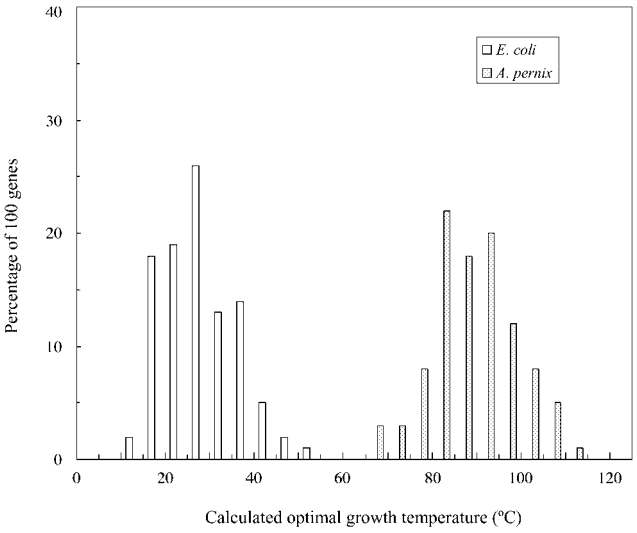

Fig. 2. **Temperature distributions of *E. coli* and *A. pernix* genes.** The temperature was calculated with Equation (2) for each of 100 genes of *E. coli* and *A. pernix* (as representatives of mesophiles and thermophiles, respectively). See Materials and Methods for the choice of the 100 genes from each bacterial species.

Figure 1 (bottom panel) is a plot of OGT-AA using the average amino acid composition of each species versus real OGT (correlation coefficient: 0.94). In this case, too, the psychrophile data deviated upward from the linear relation. The correlation coefficient dropped to 0.79 when the data for all 2,223 sequences were individually utilized. In Eq. 3, hydrophobic residues such as Ile, Val, and Leu, and charged residues such as Glu, Arg and Lys have positive coefficients, while other residues such as Gln, Asp, Asn, Ser, Thr, Cys, His, and Ala possess negative coefficients. This is consistent with the previous studies (*8*, *9*) showing that proteins in thermophiles are rich in both hydrophobic and charged residues compared to those of mesophiles.

When Eq. 3 was applied to individual proteins of *E. coli* (mesophile) and *A. pernix* (thermophile) to make histograms, the distributions of proteins from the two species were not as well-separated as in Fig. 2: some overlapping between the two was evident. This tendency is also detectable in Fig. 1, where the temperatures estimated as OGT-AA range from 30 to 70°C (see the ordinate of the bottom panel) whereas those in the case of OGT-DNA vary between 20 and 90°C (middle panel). These results imply that, in contrast to the DNA dinucleotide composition, the amino acid composition is not a precise enough measure for accurately obtaining absolute OGT values, despite its linear relationship with real OGT. The reason for this difference is discussed in the next section.

DISCUSSION

For nucleic acids, a simple way to attain thermostability is to increase the G+C content. This is indeed the case for the 16S rRNAs of thermophiles, in which the G+C content is proportional to actual OGT. If DNA adopted the same strategy for thermostability as RNA, the increased G+C content would have significant effects on the amino acid composition. For example, amino acids encoded by G+C rich codons such as Ala, Arg, Gly, and Pro would be abundant, while those encoded by G+C poor codons such as Lys, Ile, Tyr, and Phe would be less represented in thermophilic proteins. In reality, however, no correlation was observed between the G+C content of DNA and actual OGT. Instead, the dinucleotide composition of DNA was found to correlate with the adaptation to higher temperatures. Adaptation through alterations in dinucleotides would be better than that through changes in mononucleotides (G+C content), because the former approach allows more freedom as to the production of various DNA sequences encoding the amino acids of proteins.

The dinucleotide composition is known to account for the curvature (*18*) and therefore the supercoiling of double-stranded DNA. For examination of this physicochemical aspect of DNA regarding thermostability, the following index has been introduced:

$$\text{Index} = \Sigma f_i \, p_i \qquad (4)$$

where f_i is the fractional frequency of dinucleotide i and p_i is a physicochemical parameter of DNA for dinucleotide i. Note that no adjustable parameters are included in this definition. Correlations between actual OGT and the index were examined for available dinucleotide physicochemical parameters, such as the twist angle and roll angle. With the use of the dinucleotide composition data for 2,223 DNA sequences, a significant correlation with actual OGT was found in the difference in free energies of DNA base stacking (*19*) (correlation coefficient: 0.75) and in the twist angle (*20*) (correlation coefficient: –0.70). Although the index defined by Equation (4) implies some synthetic propensity and does not represent exact physical quantities, the results suggest that the supercoiling of double-stranded DNA, which is determined by the dinucleotide composition, actually affects the thermostability of DNA. Although no experimental data are available at present, it would be interesting to determine how different the shapes of genomic DNAs are between mesophilic and thermophilic bacteria.

Many reports have shown that proteins in thermophiles have, on average, different amino acid compositions from those in mesophiles (*8*, *9*). The present study further indicated that the deviation in the amino acid composition is almost linearly correlated with the real OGT of thermophilic bacteria (Fig. 1). In our previous study (*10*), the most prominent feature of thermophilic proteins was the relative abundance of charged residues (particularly, Glu, Arg and Lys) on the molecular surface, while no specific features of the composition were observed in the interior of proteins. This fact may account for the inferiority of OGT-AA in comparison with OGT-DNA as an OGT predictor, as pointed out under Results, because the amino acid composition of whole proteins instead of that of the surface was employed in the present analysis. However, estimation of the surface composition is not straightforward because it requires knowledge of the 3D structures of individual proteins.

Taking everything into account, the present study has revealed that the stabilization of genomic DNA is directly related with the dinucleotide composition, while protein

stability is indirectly related to the amino acid composition through the 3D structure because only the surface composition may be relevant. The latter conclusion should be confirmed by future studies. The general features for DNA and protein stabilization described above prompt the following conjecture: all proteins as well as genes must simultaneously acquire thermostability in thermophilic bacteria. A simple but general method for satisfying this requirement is to bias the (di)nucleotide composition of DNA in such a direction in the (di)nucleotide-composition space (*21*) as to cause an increase in charged residues at the protein level. Of course, there must be some fine-tuning in the process, such as to increase charged residues only on the protein surface. However, this simple basic strategy seems adoptable to fulfill the compositional requirements of the thermostability of both DNA and proteins at the same time.

In the present study, a limited amount of sequence data for psychrophiles was also included. The data points for psychrophiles somewhat deviated from the regression lines for RNA, DNA and proteins (Fig. 1). It is unclear from the present results whether the underlying strategy of psychrophiles is just the reverse of that used by thermophiles. As no complete genomic data for any psychrophiles are available at this moment, we must await the accumulation of more sequence data in order to elucidate the mechanism adopted by psychrophiles.

We are grateful to Dr. Keiichi Homma for the valuable comments. This work was supported in part by a Grant-in-Aid from the Ministry of Education, Culture, Sports, Science and Technology of Japan, and also supported by BIRD of Japan Science and Technology Corporation (JST).

REFERENCES

1. Galtier, N. and Lobry, J.R. (1997) Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. *J. Mol. Evol.* **44**, 632–636
2. Galtier, N., Tourasse, N., and Gouy, M. (1999) A nonhyperthermophilic common ancestor to extant life forms. *Science* **283**, 220–221
3. Hurst, L.D. and Merchant, A.R. (2001) High guanine-cytosine content is not an adaptation to high temperature: a comparative analysis amongst prokaryotes. *Proc. R. Soc. Lond. B. Biol. Sci.* **268**, 493–497
4. Lynn, D.J., Singer, G.A.C., and Hickey, D.A. (2002) Synonymous codon usage is subject to selection in thermophilic bacteria. *Nucleic Acids Res.* **30**, 4272–4277
5. Kawashima, T., Amano, N., Koike, H., Makino, S., Higuchi, S., Kawashima-Ohya, Y., Watanabe, K., Yamazaki, M., Kanehori, K., Kawamoto, T., Nunoshiba, T., Yamamoto, Y., Aramaki, H., Makino, K., and Suzuki, M. (2000) Archaeal adaptation to higher temperatures revealed by genomic sequence of *Thermoplasma volcanium*. *Proc. Natl Acad. Sci. USA* **97**, 14257–14262
6. Perutz, M.F. and Raidt, H. (1975) Stereochemical basis of heat stability in bacterial ferredoxins and in haemoglobin A2. *Nature* **255**, 256–259
7. Vogt, G., Woell, S., and Argos, P. (1977) Protein thermal stability, hydrogen bonds, and ion pairs. *J. Mol. Biol.* **269**, 631–643
8. Thompson, M.J. and Eisenberg, D. (1999) Transproteomic evidence of a loop-deletion mechanism for enhancing protein thermostability. *J. Mol. Biol.* **290**, 595–604
9. Kreil, D.P. and Ouzounis, C.A. (2001) Identification of thermophilic species by the amino acid compositions deduced from their genomes. *Nucleic Acids Res.* **29**, 1608–1615
10. Fukuchi, S. and Nishikawa, K. (2001) Protein surface amino acid compositions distinctively differ between thermophilic and mesophilic bacteria. *J. Mol. Biol.* **309**, 835–843
11. Chakravarty, S. and Varadarajan, R. (2002) Elucidation of factors responsible for enhanced thermal stability of proteins: a structural genomics based study. *Biochemistry* **41**, 8152–8161
12. Kawabata, T., Fukuchi, S., Homma, K., Ota, M., Araki, J., Ito, T., Ichiyoshi, N., and Nishikawa, K. (2002) GTOP: a database of protein structures predicted from genome sequences. *Nucleic Acids Res.* **30**, 294–298
13. Hirokawa, T., Boon-Chieng, S., and Mitaku, S. (1998) SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinformatics* **14**, 378–379
14. Nielsen, H., Brunak, S., and von Heijne, G. (1999) Machine learning approaches for the prediction of signal peptides and other protein sorting signals. *Protein Eng.* **12**, 3–9
15. Bairoch, A. and Apweiler, R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28**, 45–48
16. Tateno, Y. and Gojobori T. (1997) DNA Data Bank of Japan in the age of information biology. *Nucleic Acids Res.* **25**, 14–17
17. Muto, A. and Osawa, S. (1987) The guanine and cytosine content of genomic DNA and bacterial evolution. *Proc. Natl Acad. Sci. USA* **84**, 166–169
18. Bolshoy, A., McNamara, P., Harrington, R.E., and Trifonov, E.N. (1991) Curved DNA without A-A: experimental estimation of all 16 DNA wedge angles. *Proc. Natl Acad. Sci. USA* **88**, 2312–2316
19. Delcourt, S.G. and Blake, R.D. (1991) Stacking energies in DNA. *J. Biol. Chem.* **266**, 15160–15169
20. Kabsch, W., Sander, C., and Trifonov, E.N. (1982) The ten helical twist angles of B-DNA. *Nucleic Acids Res.* **10**, 1097–1104
21. Nakashima, H., Ota, M., Nishikawa, K., and Ooi, T. (1998) Gene from nine genomes are separated into their organisms in the dinucleotide composition space. *DNA Res.* **5**, 251–259
22. 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., Weidman, J.F., Fuhrmann, J.L., Nguyen, D., Utterback, T.R., Kelley, J.M., Peterson, J.D., Sadow, P.W., Hanna, M.C., Cotton, M.D., Roberts, K.M., Hurst, M.A., Kaine, B.P., Borodovsky, M., Klenk, H.-P., Fraser, C.M., Smith, H.O., Woese, C.R., and Venter, J.C. (1996) Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science* **273**, 1058–1073
23. Kawarabayasi, Y., Hino, Y., Horikawa, H., Jin-no, K., Takahashi, M., Sekine, M., Baba, S., Ankai, A., Kosugi, H., Hosoyama, A., Fukui, S., Nagai, Y., Nishijima, K., Otsuka, R., Nakazawa, H., Takamiya, M., Kato, Y., Yoshizawa, T., Tanaka, T., Kudoh, Y., Yamazaki, J., Kushida, N., Oguchi, A., Aoki, K., Masuda, S., Yanagii, M., Nishimura, M., Yamagishi, A., Oshima, T. and Kikuchi, H. (2001) Complete genome sequence of an aerobic thermoacidophilic crenarchaeon, *Sulfolobus tokodaii* strain 7. *DNA Res.* **8**, 123–140
24. She, Q., Singh, R.K., Confalonieri, F., Zivanovic, Y., Allard, G., Awayez, M.J., Chan-Weiher, C.C.-Y., Clausen, I.G., Curtis, B.A., De Moors, A., Erauso, G., Fletcher, C., Gordon, P.M.K., Jong, I.H., Jeffries, A.C., Kozera, C.J., Medina, N., Peng, X., Thi-Ngoc, H.P., Redder, P., Schenk, M.E., Theriault, C., Tolstrup, N., Charlebois, R.L., Doolittle, W.F., Duguet, M., Gaasterland, T., Garrett, R.A., Ragan, M.A., Sensen, C.W., and Van der Oost, J. (2001) The complete genome of the crenarchaeon *Sulfolobus solfataricus* P2. *Proc. Natl Acad. Sci. USA* **98**, 7835–7840
25. Kawashima, T., Yamamoto, Y., Aramaki, H., Nunoshiba, T., Kawamoto, T., Watanabe, K., Yamazaki, M., Kanehori, K., Amano, N., Ohya, Y., Makino, K., and Suzuki, M. (1999) Determination of the complete genomic DNA sequence of *Thermoplasma volcanium* GSS1. *Proc. Jpn. Acad.* **75**B, 213–218

26. 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., Otuka, R., Nakazawa, H., Takamiya, M., Ohfuku, Y., Funahashi, T., Tanaka, T., Kudoh, Y., Yamazaki, J., Kushida, N., Oguchi, A., Aoki, K., Nakamura, Y., Robb, T.F., Horikoshi, K., Masuchi, Y., Shizuya, H. and Kikuchi, H. (1998) Complete sequence and gene organization of the genome of a hyper-thermophilic archaebacterium, *Pyrococcus horikoshii* OT3. *DNA Res.* **5**, 55–76
27. Ruepp, A., Graml, W., Santos-Martinez, M.L., Koretke, K.K., Volker, C., Mewes, H.W., Frishman, D., Stocker, S., Lupas, A.N., and Baumeister, W. (2000) The genome sequence of the thermoacidophilic scavenger *Thermoplasma acidophilum*. *Nature* **407**, 508–513
28. 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., Peterson, S., Reich, C.I., McNeil, L.K., Badger, J.H., Glodek, A., Zhou, L., Overbeek, R., Gocayne, J.D., Weidman, J.F., McDonald, L., Utterback, T., Cotton, M.D., Spriggs, T., Artiach, P., Kaine, B.P., Sykes, S.M., Sadow, P.W., D'Andrea, K.P., Bowman, C., Fujii, C., Garland, S.A., Mason, T.M., Olsen, G.J., Fraser, C.M., Smith, H.O., Woese, C.R., and Venter, J.C. (1997) The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. *Nature* **390**, 364–370
29. Smith, D.R., Doucette-Stamm, L.A., Deloughery, C., Lee, H.-M., 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., Vicare, R., Wang, Y., Wierzbowski, J., Gibson, R., Jiwani, N., Caruso, A., Bush, D., Safer, H., Patwell, D., Prabhakar, S., McDougall, S., Shimer, G., Goyal, A., Pietrovski, S., Church, G.M., Daniels, C.J., Mao, J.-i., Rice, P., Nolling, J., and Reeve, J.N. (1997) Complete genome sequence of *Methanobacterium thermoautotrophicum* delta H: functional analysis and comparative genomics. *J. Bacteriol.* **179**, 7135–7155
30. Kawarabayasi, Y., Hino, Y., Horikawa, H., Yamazaki, S., Haikawa, Y., Jin-no, K., Takahashi, M., Sekine, M., Baba, S., Ankai, A., Kosugi, H., Hosoyama, A., Fukui, S., Nagai, Y., Nishijima, K., Nakazawa, H., Takamiya, M., Masuda, S., Funahashi, T., Tanaka, T., Kudoh, Y., Yamazaki, J., Kushida, N., Oguchi, A., Aoki, K., Kubota, K., Nakamura, Y., Nomura, N., Sako, Y., and Kikuchi, H. (1999) Complete genome sequence of an aerobic hyper-thermophilic crenarchaeon, *Aeropyrum pernix* K1. *DNA Res.* **6**, 83–101
31. 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., Olson, G.J., and Swanson, R.V. (1998) The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. *Nature* **392**, 353–358
32. Nelson, K.E., Clayton, R.A., Gill, S.R., Gwinn, M.L., Dodson, R.J., Haft, D.H., Hickey, E.K., Peterson, J.D., Nelson, W.C., Ketchum, K.A., McDonald, L., Utterback, T.R., Malek, J.A., Linher, K.D., Garrett, M.M., Stewart, A.M., Cotton, M.D., Pratt, M.S., Phillips, C.A., Richardson, D., Heidelberg, J., Sutton, G.G., Fleischmann, R.D., Eisen, J.A., White, O., Salzberg, S.L., Smith, H.O., Venter, J.C., and Fraser, C.M. (1999) Evidence for lateral gene transfer between Archaea and Bacteria from genome sequence of *Thermotoga maritime*. *Nature* **399**, 323–329
33. Parkhill, J., Wren, B.W., Mungall, K., Ketley, J.M., Churcher, C., Basham, D., Chillingworth, T., Davies, R.M., Feltwell, T., Holroyd, S., Jagels, K., Karlyshev, A.V., Moule, S., Pallen, M.J., Penn, C.W., Quail, M.A., Rajandream, M.-A., Rutherford, K.M., Van Vliet, A.H.M., Whitehead, S., and Barrell, B.G. (2000) The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. *Nature* **403**, 665–668
34. Fraser, C.M., Gocayne, J.D., White, O., Adams, M.D., Clayton, R.A., Fleischmann, R.D., Bult, C.J., Kerlavage, A.R., Sutton, G., Kelley, J.M., Fritchman, J.L., Weidman, J.F., Small, K.V., Sandusky, M., Fuhrmann, J., Nguyen, D., Utterback, T.R., Saudek, D.M., Phillips, C.A., Merrick, J.M., Tomb, J.-F., Dougherty, B.A., Bott, K.F., Hu, P.-C., Lucier, T.S., Peterson, S.N., Smith, H.O., Hutchison, C.A. III, and Venter, J.C. (1995) The minimal gene complement of *Mycoplasma genitalium*. *Science* **270**, 397–403
35. 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., Mckenney, K., Sutton, G., FitzHugh, W., Fields, C., Gocayne, J.D., Scott, J., Shirley, R., Liu, L.-I., Glodek, A., Kelley, J.M., Weidman, J.F., Philips, C.A., Spriggs, T., Hedbolm, E., Cotton, M.D., Utterback, T.R., Hanna, M.C., Nguyen, D.T., Saudek, D.M., Brandon, R.C., Fine, L.D., Fritchman, J.L., Fuhrmann, J.L., Geoghagen, N.S.M., Gnehm, C.L., McDonald, L.A., Small, K.V., Fraser, C.M., Smith, H.O., and Venter, J.C. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* rd. *Science* **269**, 496–512
36. 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., Hickey, E.K., Berg, D.E., Gocayne, J.D., Utterback, T.R., Peterson, J.D., Kelley, J.M., Cotton, M.D., Weidman, J.M., Fujii, C., Bowman, C., Watthey, L., Wallin, E., Hayes, W.S., Borodovsky, M., Karp, P.D., Smith, H.O., Fraser C.M., and Venter, J.C. (1997) The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* **388**, 539–547
37. Himmelreich, R., Hilbert, H., Plagens, H., Pirkl, E., Li, B.-C., and Herrmann, R. (1996) Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucleic Acids Res.* **24**, 4420–4449
38. Kunst, F., Ogasawara, N., Moszer, I., Albertini, A.M., Alloni, G., Azevedo, V., Bertero, M.G., Bessieres, P., Bolotin, A., Borchert, S., Borriss, R., Boursier, L., Brans, A., Braun, M., Brignell, S.C., Bron, S., Brouillet, S., Bruschi, C.V., Caldwell, B., Capuano, V., Carter, N.M., Choi, S.K., Codani, J.J., Connerton, I.F., Cummings, N.J., Daniel, R.A., Denizot, F., Devine, K.M., Dusterhoft, A., Ehrlich, S.D., Emmerson, P.T., Entian, K.D., Errington, J., Fabret, C., Ferrari, E., Foulger, D., Fritz, C., Fujita, M., Fujita, Y., Fuma, S., Galizzi, A., Galleron, N., Ghim, S.Y., Glaser, P., Goffeau, A., Golightly, E.J., Grandi, G., Guiseppi, G., Guy, B.J., Haga, K., Haiech, J., Harwood, C.R., Henaut, A., Hilbert, H., Holsappel, S., Hosono, S., Hullo, M.F., Itaya, M., Jones, L., Joris, B., Karamata, D., Kasahara, Y., Klaerr-Blanchard, M., Klein, C., Kobayashi, Y., Koetter, P., Koningstein, G., Krogh, S., Kumano, M., Kurita, K., Lapidus, A., Lardinois, S., Lauber, J., Lazarevic, V., Lee, S.M., Levine, A., Liu, H., Masuda, S., Mauel, C., Medigue, C., Medina, N., Mellado, R.P., Mizuno, M., Moestl, D., Nakai, S., Noback, M., Noone, D., O'Reilly, M., Ogawa, K., Ogiwara, A., Oudega, B., Park, S.H., Parro, V., Pohl, T.M., Portetelle, D., Porwollik, S., Prescott, A.M., Presecan, E., Pujic, P., Purnelle, B., Rapoport, G., Rey, M., Reynolds, S., Rieger, M., Rivolta, C., Rocha, E., Roche, B., Rose, M., Sadaie, Y., Sato, T., Scanlan, E., Schleich, S., Schroeter, R., Scoffone, F., Sekiguchi, J., Sekowska, A., Seror, S.J., Serror, P., Shin, B.S., Soldo, B., Sorokin, A., Tacconi, E., Takagi, T., Takahashi, H., Takemaru, K., Takeuchi, M., Tamakoshi, A., Tanaka, T., Terpstra, P., Tognoni, A., Tosato, V., Uchiyama, S., Vandenbol, M., Vannier, F., Vassarotti, A., Viari, A., Wambutt, R., Wedler, E., Wedler, H., Weitzenegger, T., Winters, P., Wipat, A., Yamamoto, H., Yamane, K., Yasumoto, K., Yata, K., Yoshida, K., Yoshikawa, H.F., Zumstein, E., Yoshikawa, H., and Danchin, A. (1997) The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. *Nature* **390**, 249–256
39. Parkhill, J., Wren, B.W., Thomson, N.R., Titball, R.W., Holden, M.T.G., Prentice, M.B., Sebaihia, M., James, K.D., Churcher, C., Mungall, K.L., Baker, S., Basham, D., Bentley, S.D., Brooks, K., Cerdeno-Tarraga, A.M., Chillingworth, T., Cronin, A., Davies, R.M., Davis, P., Dougan, G., Feltwell, T., Hamlin, N., Hol-

royd, S., Jagels, K., Karlyshev, A.V., Leather, S., Moule, S., Oyston, P.C.F., Quail, M., Rutherford, K., Simmonds, M., Skelton, J., Stevens, K., Whitehead, S., and Barrell, B.G. (2001) Genome sequence of *Yersinia pestis*, the causative agent of plague. *Nature* **413**, 523–527

40. Blattner, F.R., Plunkett, G. III, 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., and Shao, Y. (1997) The complete genome sequence of *Escherichia coli* K-12. *Science* **277**, 1453–1462
41. Cole, S.T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S.V., Eiglmeier, K., Gas, S., Barry III, C.E., Tekaia, F., Badcock, K., Basham, D., Brown, D., Chillingworth, T., Connor, R., Davies, R., Devlin, K., Feltwell, T., Gentles, S., Hamlin, N., Holroyd, S., Hornsby, T., Jagels, K., Krogh, A., McLean, J., Moule, S., Murphy, L., Oliver, S., Osborne, J., Quail, M.A., Rajandream, M.A., Rogers, J., Rutter, S., Seeger, K., Skelton, S., Squares, S., Sqares, R., Sulston, J.E., Taylor, K., Whitehead, S., and Barrell, B.G. (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393**, 537–544
42. Stover, C.K., Pham, X.Q., Erwin, A.L., Mizoguchi, S.D., Warrener, P., Hickey, M.J., Brinkman, F.S.L., Hufnagle, W.O., Kowalik, D.J., Lagrou, M., Garber, R.L., Goltry, L., Tolentino, E., Westbrock-Wadman, S., Yuan, Y., Brody, L.L., Coulter, S.N., Folger, K.R., Kas, A., Larbig, K., Lim, R., Smith, K., Spencer, D., Wong, G.K.-S., Wu, Z., Paulsen, I.T., Reizer, J., Saier, M.H., Hancock, R.E.W., Lory, S., and Olson, M.V. (2000) Complete genome sequence of *Pseudomonas aeruginosa* PA01, an opportunistic pathogen. *Nature* **406**, 959–964
43. Preston, C.M., Wu, K.Y., Molinski, T.F., and DeLong, E.F. (1996) A psychrophilic crenarchaeon inhabits a marine sponge: *Cenarchaeum symbiosum* gen. nov., sp. nov. *Proc. Natl Acad. Sci. USA* **93**, 6241–6246