Comparative Analysis of DNA-Binding Proteins between Thermophilic and Mesophilic Bacteria

Masashi Fujita fujita@kuicr.kyoto-u.ac.jp
Minoru Kanehisa kanehisa@kuicr.kyoto-u.ac.jp

Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011, Japan

Abstract

Thermophilic bacteria are one of the most attractive forms of life, and their adaptation mechanisms to elevated temperatures have been extensively studied over the years. Thermal adaptations of cell components such as proteins and RNA are well studied, but adaptations of interactions between these components must be also vital for the thermophiles. Protein-DNA interactions play crucial roles in the cell, but little is known about their thermal adaptations. In this study, we analyzed DNA-binding proteins from thermophilic bacteria. Comparison of amino acid compositions at the DNA-binding interfaces between thermophiles and their mesophilic close relatives revealed several commonalities between phylogenetically unrelated organisms. Advantages and limitations of our methods will be also discussed.

Keywords: thermal adaptation, DNA-binding protein, amino acid composition

1 Introduction

Adaptations of thermophilic microorganisms to high environmental temperature have been under extensive study for years [5, 9]. These efforts have provided valuable insights into the thermophile's biology and clarified how various components of the cell have adapted to their extreme living conditions. This includes not only DNA, RNA and proteins but also other small molecules such as membrane lipids.

The best studied cell components of thermophiles are proteins [9]. Proteins from thermophiles have characteristic amino acid compositions compared with mesophiles [3, 14]. That is mainly because of the differences in composition on the protein surface [3]. The surface regions of thermophilic proteins have fewer (non-charged) polar amino acids and more charged amino acids, and these charged residues result in an increased number of intramolecular salt bridges. Decreased sequence length is another feature of thermophilic proteins [17]. They tend to be shorter than their mesophilic homologs, and the main cause of this length reduction is the deletions in the loop regions.

Thermal adaptations of RNA molecules are also known. rRNA and tRNA molecules of thermophilic bacteria have higher G+C contents than mesophiles [4]. Because the GC base pair forms more hydrogen bonds than the AT base pair, higher G+C contents in the double-stranded stem region improves thermostability of the RNA molecules. Additionally, mRNA of thermophiles contain more purine nucleotides [10, 12]. A likely explanation of this phenomenon is that it is to prevent aggregation of mRNA molecules [10]. Adaptation of genomic DNA is not as simple as RNA. There is no apparent correlation between the G+C contents of the DNA and the optimal growth temperature of the organism [4]. Instead, several thermophiles have special proteins that bind to DNA and raise its melting temperature significantly. Furthermore, the cell membrane of thermophiles contain an increased ratio of saturated lipid and maintain their fluidity.

As made apparent above, each cell component has distinctive characteristics to stabilize itself and tolerate elevated temperatures. But the cell is a complicated system; not only do its components

but also the interactions between them need to have enhanced stability. Protein-DNA interactions play crucial roles in the living cell. A number of vital processes such as chromosomal replication, transcriptional regulation and mismatch repair require proper protein-DNA interactions. Therefore, if protein-DNA interactions become unstable at elevated temperatures, the fitness of the organism will decrease. This leads us to a simple hypothesis. Either the DNA-binding proteins and/or their binding sites on the DNA of thermophilic organisms will be adapted to high temperatures to manage thermal instability.

Although this hypothesis seems plausible, little is known about the thermal adaptations of protein-DNA interactions. Hopfner *et al.* reported that DNA polymerase from hyperthermophilic Archaea displays an enhanced electrostatic complementarity at the DNA-protein interface [6]. But no comprehensive analyses on this problem have been presented to date. In recent years, several thermophilic Bacterial and Archaeal genomes have been decoded, and these invaluable resources could shed light on the above problem. The adaptations of protein-binding sites on the DNA are still difficult to analyze even though the genome is available. This is because most of the binding sites are as of yet experimentally undiscovered, and computational methods for *cis*-element inference remain immature. In contrast, an analysis of the DNA-binding proteins is feasible. DNA-binding proteins can be identified by sequence homology, and their binding interfaces can be investigated based on the 3D coordinates.

Here we collected DNA-binding proteins from several thermophilic bacteria and compared their DNA-binding interfaces with mesophilic orthologs. Analyses of amino acid compositions at the DNA-binding interfaces revealed several commonalities among phylogenetically-unrelated thermophilic organisms.

2 Materials and Methods

2.1 Target Organisms

The organisms studied are 12 thermophiles, 7 of which are Bacteria and 5 of which are Archaea (Table 2.1). Additionally, one or more mesophiles were selected as control organisms for each thermophile. Close relatives of the thermophile were selected as the control organisms whenever possible. Although *S. thermophilum* had been classified into Actinobacteria, its genome sequence revealed that it is closer to Firmicutes [18]. We thus selected 7 mesophilic Firmicutes as control organisms of *S. thermophilum*. Control organisms of *A. fulgidus* were selected from Methanosarcinales and Halobacteriales, because recent phylogenetic analysis confirmed their relatedness [2]. For *T. maritima*, *A. aeolicus*, *N. equitans*, *P. horikoshii* and *P. aerophilum*, appropriate mesophile genomes were not available. We had to use average of several mesophiles as a substitute.

If an organism was in the NCBI COG database [15, 16], ORF data of the organism was downloaded from NCBI's FTP site [19]. Otherwise amino acid sequences of the organism were downloaded from the KEGG GENES database [8], and putative COG IDs were assigned to them by emulating COGNITOR, an automatic COG assignment system [15].

2.2 Selection of DNA-Binding Proteins

X-ray crystal structures of protein-DNA complexes whose resolution is better than 3.0 Å were downloaded from PDB [1] on 9 January 2005. If all the DNA chains in a structure were shorter than eight nucleotides, the structure was discarded. Peptide chains longer than 40 residues were extracted from the structures. This resulted in 1349 peptide chains from 653 structures. To remove redundancy, these proteins were clustered at 90 % sequence identity and representatives were selected using CD-HIT [11]. 250 representative proteins were thus selected, and COG IDs were assigned to them using the COGNITOR-like procedure. After ID assignment, Bacteria and Archaea were treated independently.

The Bacterial DNA-binding proteins were selected as follows. Each of 250 representative proteins

Table 1: Target organisms. The thermophilic organisms studied are described in the first column. B and A in the second column denote Bacteria and Archaea. The third column is the optimal growth temperatures of the organisms. Mesophilic organisms in the fourth column were used as control and compared with the thermophiles.

Thermophile	Domain	OGT (°C)	Control Mesophiles
Geobacillus kaustophilus	B	55	*Bacillus subtilis* & *B. halodurans*
Thermosynechococcus elongatus	B	55	*Synechocystis* & *Anabaena*
Symbiobacterium thermophilum	B	60	7 mesophilic Firmicutes
Thermoanaerobacter tengcongensis	B	75	*Clostridium acetobutylicum*
Thermus thermophilus HB27	B	80	*Deinococcus radiodurans*
Thermotoga maritima	B	80	38 mesophilic Bacteria
Aquifex aeolicus	B	85	38 mesophilic Bacteria
Methanococcus jannaschii	A	85	*Methanococcus maripaludis*
Archaeoglobus fulgidus	A	85	Methanosarcinales & Halobacteriales
Nanoarchaeum equitans	A	90	6 mesophilic Euryarchaeaota
Pyrococcus horikoshii	A	95	6 mesophilic Euryarchaeaota
Pyrobaculum aerophilum	A	98	6 mesophilic Euryarchaeaota

was examined to determine whether its assigned COG ID was "common" or not in the Bacterial proteome. Specifically, the COG ID was required to be found in at least three thermophilic bacteria and three mesophilic bacteria. If this condition was not satisfied, the protein was excluded from subsequent analyses. The remaining proteins were manually investigated, and inappropriate proteins such as RNA-binding proteins or peptide chains that have no direct contact with DNA were removed. The proteins were further clustered into families based on the COG IDs. As a result, 28 Bacterial DNA-binding protein families were selected. The same procedure was applied to Archaea, and 22 families were obtained. For each of these families, multiple structural alignment of its members was calculated using STAMP [13], and a profile hidden Markov model for HMMER [20] was constructed.

2.3 Amino Acid Compositions at the DNA-Binding Interfaces

In this study, we simply define the DNA-binding interface of a protein as the set of residues in contact with DNA (within 3.8 Å). Because the above HMM models were derived from multiple alignments of proteins with known structures, the interface residues can be mapped to matching states of the HMM models. If a protein from bacteria belonged to the same COG with one of the DNA-binding protein families, its sequence was aligned to the HMM model and its DNA-binding interface was inferred from the alignment.

To reduce bias from differences in member sizes and sequence length between protein families, an average amino acid composition at the binding interfaces of an organism was calculated assigning each protein family equal weight.

3 Results and Discussion

3.1 Strategy

If DNA-binding proteins from thermophilic microorganisms are adapted to high environmental temperatures, this would be most reflected in their DNA-binding interfaces. Detailed information on the interfaces can be obtained from 3D structures of the protein-DNA complexes. But solved structures of the protein-DNA complexes are still scarce compared with recent accumulation of sequence information. Thus, we exploited sequence homology, and inferred DNA-binding interfaces of proteins

from sequence alignments with the structures of the protein-DNA complexes. Because amino acid composition allows us to understand global features of proteins, average amino acid compositions at the DNA-binding interfaces were calculated for each thermophilic bacteria.

It is obvious that the binding interfaces of thermophilic bacteria should be compared with that of mesophilic bacteria. But selecting the organisms to be compared is not a trivial task. Although it is tempting to analyze all available thermophile genomes, such analyses may suffer from lack of appropriate mesophile genomes. For example, most Archaea whose genome sequences have been determined are hyperthermophilic. Several mesophilic Archaea genomes are available, but they are biased toward specific taxonomic groups, such as Methanococci. This makes straightforward comparisons between thermophilic and mesophilic Archaea problematic. Even though such comparisons may reveal differences, it is difficult to tell whether the differences reflect thermal adaptations or features of Methanococci.

Thermophilic organisms are found in diverse branches of the phylogenetic tree. This means that the adaptations to high temperatures may have occurred multiple times and independently in the course of evolution. Commonality among the independent evolutionary events would be strong evidence of thermal adaptations. Therefore, a thermophile should be compared only with its closely related mesophilic species. If these organisms have some characteristic differences, and the same trends are repeatedly observed among multiple taxonomic groups, then we can believe that the difference is a result of thermal adaptations. Although this strategy seems to be ideal for analyses of thermal adaptations, it was rarely adopted by previous researchers. This may simply be because closely related pairs of thermophile and mesophile genomes are few. In this study, we followed this strategy when possible.

3.2 Amino Acid Composition

Average amino acid compositions at the DNA-binding interfaces of thermophilic Bacteria, mesophilic Bacteria, thermophilic Archaea and mesophilic Archaea are shown (Figure 1). The distribution of these four groups are largely similar. High contents of positively charged residues (Arg and Lys) support the importance of electrostatic interactions in protein-DNA interactions. Because Ser and Thr can act as both donors and acceptors of hydrogen bonds, these residues are also frequent. The difference between Bacteria and Archaea looks greater than that between mesophiles and thermophiles. This can be attributed to the fact that DNA-binding protein families used to calculate the amino acid compositions were different between Bacteria and Archaea.

Next, the individual thermophilic organisms were analyzed, and their amino acid compositions at the DNA-binding interfaces were compared with that of their mesophilic close relatives. Figure 2 shows the increase and decrease of each amino acid at the interfaces. Although each organism has a unique pattern of variation and no row is coherent, several trends can be identified. Arg is increased in general at the interfaces in all organisms except *A. aeolicus*, and Val and His are also generally increased in all but two organisms. Asn is decreased in all organisms except *M. jannaschii*. However, physicochemical interpretation of these results is somewhat unclear. It is natural to expect the total charge at the DNA-binding interfaces of thermophiles to be more positive than that of mesophiles. The increase of Arg encourages this hypothesis. But the increase of Lys, another positively charged residue, is less common (7 out of 12), and negatively charged Glu residue is also increased (8 out of 12). In fact, the total charge at the interfaces have little correlation with the optimal growth temperature of the organisms (data not shown). Hydrogen bonding is another important factor in protein-DNA interactions, but non-charged hydrogen bonding residues, such as Ser, Thr, Tyr and Asn tend to decrease.

Although the analysis of amino acid compositions did not provide us an easy-to-understand picture of the thermal adaptations, an unexpected result, the increase of His, was observed. Because most His residues are not protonated in physiological conditions, it may not be involved in electrostatic inter-

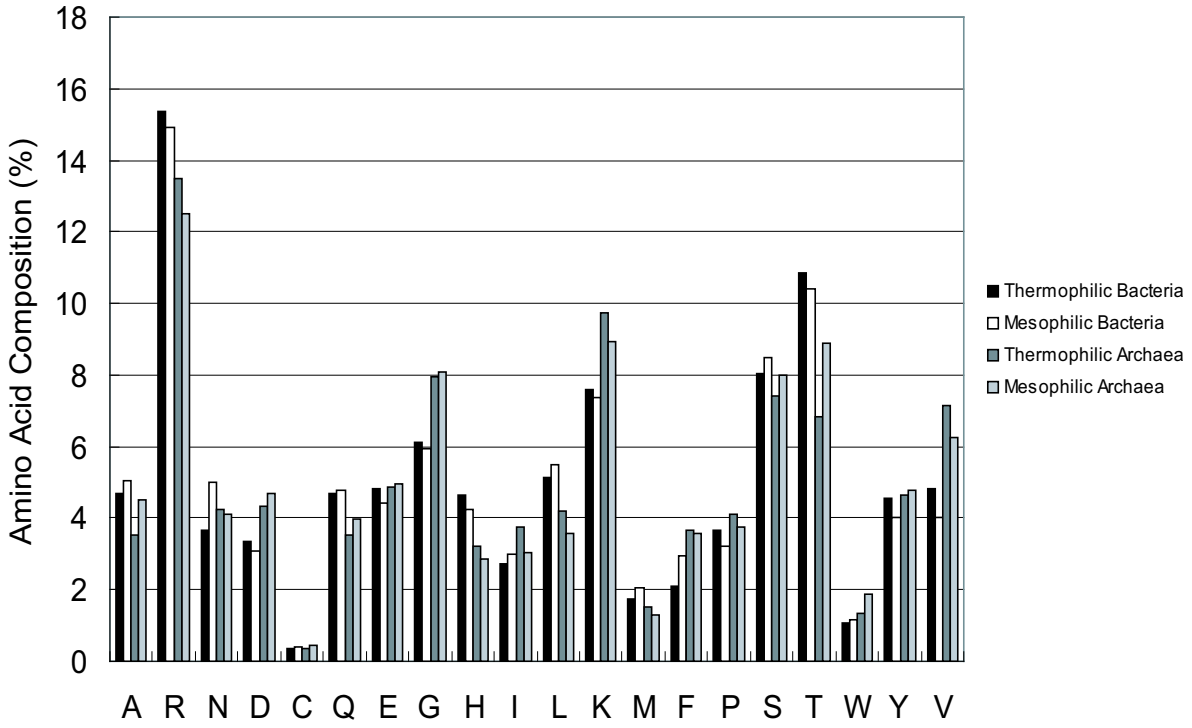

Figure 1: Average amino acid compositions at the DNA-binding interfaces of four organism groups.

actions. Moreover, His is a less favored amino acid at generic DNA-binding interfaces [7]. Therefore, the general increase of His is puzzling.

Before drawing any conclusions, not only amino acid compositions at the interfaces but also that of the whole proteome should be taken into account because the observed composition change at the interfaces could be a secondary effect of composition change in the whole proteome. We calculated the amino acid compositions of the whole proteome and compared them between thermophiles and mesophiles (Table 2). The most interesting amino acids are residues that increased at the DNA-binding interfaces but decreased in the proteome. His has such patterns in 8 out of 12 organisms. Arg is increased at the interfaces but also increased in the whole proteome. Thus it is difficult to determine whether the increased content of Arg is due to adaptations of protein-DNA interactions or that of unbound free proteins.

4 Conclusion

Thermal adaptations of DNA-binding proteins were analyzed in this paper. We adapted a rigorous strategy which analyzes phylogenetically independent events, and we compared the amino acid compositions at the DNA-binding interfaces between thermophiles and mesophiles. This revealed several commonalities in different species, although physicochemical interpretation of the results is obscure. The most counterintuitive result is the increased composition of histidine residues, which is less favored both at generic DNA-binding interfaces and in the proteome of thermophiles.

It is difficult to distinguish the adaptations of the interactions from that of their participants. Therefore, to overcome this difficulty, we compared the interfaces with the whole proteome and focused on residues that show opposite patterns of increase and decrease. However, it seems that this approach

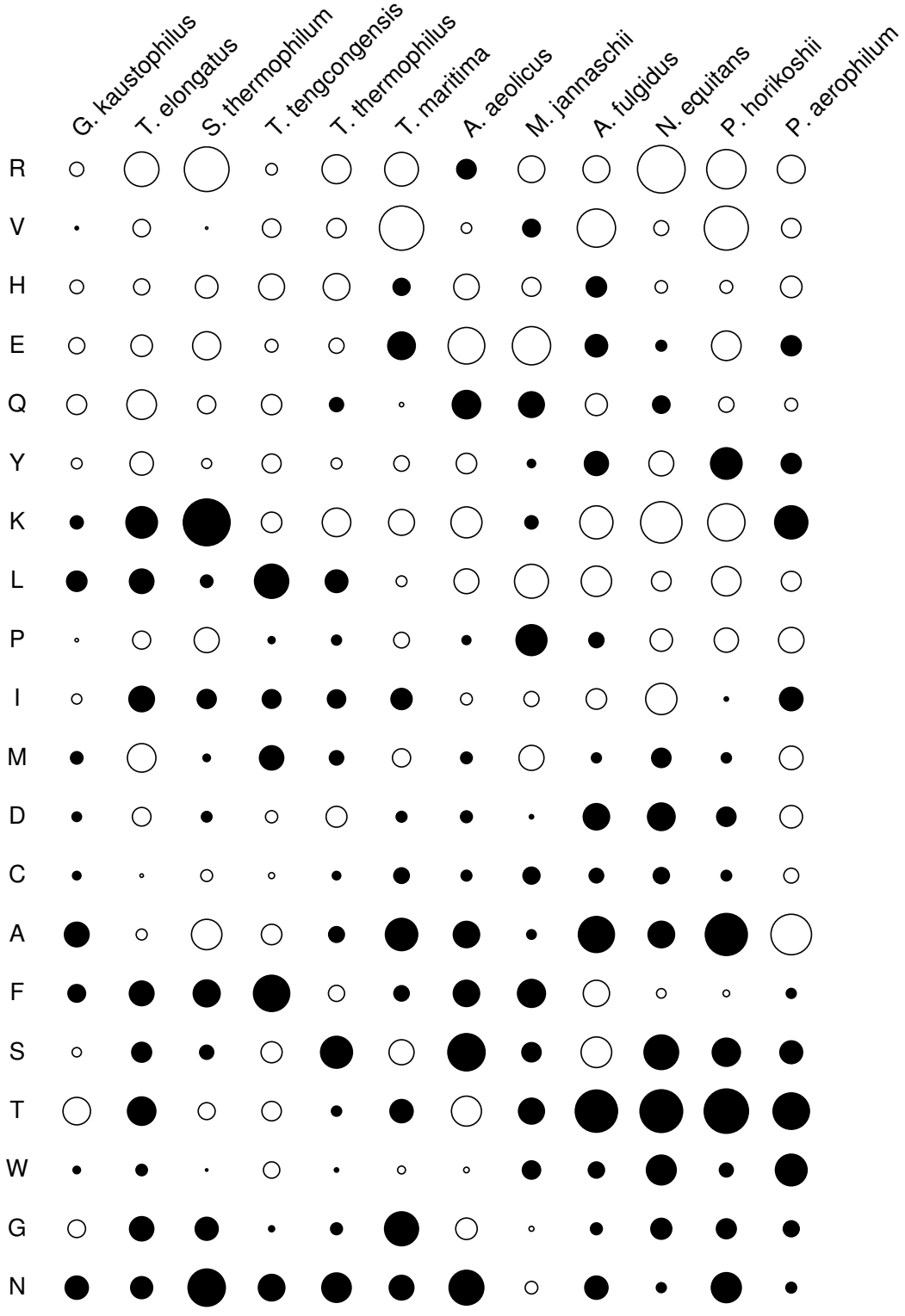

Figure 2: Increase and decrease of 20 amino acids at the DNA-binding interfaces of thermophiles compared with that of mesophiles. A white circle means the amino acid is increased in the thermophile compared with closely related mesophiles. A black circle denotes decrease. Area of a circle is proportional to the absolute value of increment or decrement.

Table 2: Comparisons of amino acid composition changes between the DNA-binding interfaces and the whole proteome. A double-plus means the amino acid is increased at the interfaces but decreased in the whole proteome. A single-plus means the amino acid is increased both at the interfaces and in the proteome, but the increment at the interfaces is greater than that in the proteome. A double-minus and a single-minus are the opposites of double-plus and single-plus.

	Gka	Tel	Sth	Tte	Tth	Tma	Aae	Mja	Afu	Neq	Pho	Pya
H	++		++	+	++		++	++		++	++	++
Q	++	+	++	+				-	++		++	++
V		+			+	+			+	++	++	
D		++		++	++							++
S				++		++			++		-	
P						++		-		++	+	+
T	++	-		++			++				-	
E	++	++	++			--		+	--		+	
R										++	+	
K				++								--
N							-	++	-		-	
Y		++		++					--		--	--
M	-	+		-	-	+		++	--	-	--	++
A	--								-		-	+
I		-		-		--			+			-
C	--		+	++	-	-	-	-	--	-		++
G		--	--			-	++					--
L	--	--	--	--	--		+	+	+		+	
F	-	-	-	-	+	--	--	--	+			
W	--	--		+		++		--	--	-	--	--

may be too conservative. More sensitive methods will be required to analyze the adaptations of intracellular interactions in future research.

Acknowledgments

This work was supported by grants from the Ministry of Education, Culture, Sports, Science and Technology, the Japan Society for the Promotion of Science, and the Japan Science and Technology Agency. The computational resource was provided by the Bioinformatics Center, Institute for Chemical Research, Kyoto University.

References

[1] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E., The Protein Data Bank, *Nucleic Acids Res.*, 28(1):235–242 2000.

[2] Brochier, C., Forterre, P., and Gribaldo, S., Archaeal phylogeny based on proteins of the transcription and translation machineries: tackling the Methanopyrus kandleri paradox, *Genome Biol.*, 5(3):R17, 2004.

[3] Fukuchi, S. and Nishikawa, K., Protein surface amino acid compositions distinctively differ between thermophilic and mesophilic bacteria, *J. Mol. Biol.*, 309(4):835–843. 2001.

[4] Galtier, N. and Lobry, J.R., Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes, *J. Mol. Evol.*, 44(6):632–636, 1997.

[5] Hickey, D.A. and Singer, G.A., Genomic and proteomic adaptations to growth at high temperature, *Genome Biol.*, 5(10):117, 2004.

[6] Hopfner, K.P., Eichinger, A., Engh, R.A., Laue, F., Ankenbauer, W., Huber, R., and Angerer, B., Crystal structure of a thermostable type B DNA polymerase from Thermococcus gorgonarius, *Proc. Natl. Acad. Sci. USA*, 96(7):3600–3605, 1999.

[7] Jones, S., Shanahan, H.P., Berman, H.M., and Thornton, J.M., Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins, *Nucleic Acids Res.*, 31(24):7189–7198, 2003.

[8] Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M., The KEGG resource for deciphering the genome, *Nucleic Acids Res.*, 32(Database issue):D277–280, 2004.

[9] Kumar, S. and Nussinov, R., How do thermophilic proteins deal with heat?, *Cell. Mol. Life Sci.*, 58(9):1216–1233, 2001.

[10] Lao, P.J. and Forsdyke, D.R., Thermophilic bacteria strictly obey Szybalski's transcription direction rule and politely purine-load RNAs with both adenine and guanine, *Genome Res.*, 10(2):228–236, 2000.

[11] Li, W., Jaroszewski, L., and Godzik, A., Tolerating some redundancy significantly speeds up clustering of large protein databases, *Bioinformatics*, 18(1):77–82, 2002.

[12] Paz, A., Mester, D., Baca, I., Nevo, E., and Korol, A., Adaptive role of increased frequency of polypurine tracts in mRNA sequences of thermophilic prokaryotes, *Proc. Natl. Acad. Sci. USA*, 101(9):2951–2956, 2004.

[13] Russell, R.B. and Barton, G.J., Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels, *Proteins*, 14(2):309–323, 1992.

[14] Singer, G.A. and Hickey, D.A., Thermophilic prokaryotes have characteristic patterns of codon usage, amino acid composition and nucleotide content, *Gene*, 317(1-2):39–47 2003.

[15] Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., Rao, B.S., Smirnov, S., Sverdlov, A.V., Vasudevan, S., Wolf, Y.I., Yin, J.J., and Natale, D.A., The COG database: an updated version includes eukaryotes, *BMC Bioinformatics*, 4(1):41, 2003.

[16] Tatusov, R.L., Koonin, E.V., and Lipman, D.J., A genomic perspective on protein families, *Science*, 278(5338):631–637, 1997.

[17] Thompson, M.J. and Eisenberg, D., Transproteomic evidence of a loop-deletion mechanism for enhancing protein thermostability, *J. Mol. Biol.*, 290(2):595–604, 1999.

[18] Ueda, K., Yamashita, A., Ishikawa, J., Shimada, M., Watsuji, T.O., Morimura, K., Ikeda, H., Hattori, M., and Beppu, T., Genome sequence of Symbiobacterium thermophilum, an uncultivable bacterium that depends on microbial commensalism, *Nucleic Acids Res.*, 32(16):4937–4944, 2004.

[19] `ftp://ftp.ncbi.nlm.nih.gov/pub/COG/`

[20] `http://hmmer.wustl.edu/`