Ann Microbiol (2012) 62:1635–1641
DOI 10.1007/s13213-012-0420-0

ORIGINAL ARTICLE

Analysis of structural requirements for thermo-adaptation from orthologs in microbial genomes

Junxiang Gao · Wei Wang

Received: 1 June 2011 / Accepted: 20 January 2012 / Published online: 9 February 2012

Abstract A comprehensive survey was carried out to identify orthologs of proteins from 526 bacterial and archaeal genomes, with the aim of investigating the mechanism of thermal adaptation of protein sequences. A large number of orthologs were distributed only in thermophiles/hyperthermophiles (HT-only group) and mesophiles (M-only group). A significant relationship between amino acid composition and optimal growth temperature (OGT) was observed. There were significantly higher proportions of charged, basic and acidic amino acids in hyperthermophilic and thermophilic genomes than in mesophilic and psychrophilic genomes. The orthologs distributed in all the four temperature ranges (Top-90 group) were also investigated, and a similar correlation between amino acid composition and OGT was found. The composition of the cluster of orthologous groups of proteins (COG) of the above three groups was analyzed; the composition of 'information storage and processing' in the HT-only group and Top-90 groups was much higher than that of M-only group.

Keywords Thermal adaptation · Bacterial and archaeal genomes · Protein sequences · Optimal growth temperature · Cluster of orthologous groups of proteins

Electronic supplementary material The online version of this article (doi:10.1007/s13213-012-0420-0) contains supplementary material, which is available to authorized users.

J. Gao (✉)
College of Science, Huazhong Agricultural University,
Wuhan 430070, People's Republic of China
e-mail: gao200@mail.hzau.edu.cn

W. Wang
Shandong Provincial Research Center for Bioinformatic
Engineering and Technique, Shandong University of Technology,
Zibo 255049, People's Republic of China

Introduction

The thermal adaptation of hyperthermophiles and their protein thermo-stability are attractive and complex topics that have drawn considerable attention (Burra et al. 2010; Dehouck et al. 2008; De Vendittis et al. 2008; Dutta and Chaudhuri 2010; Farias and Bonato 2003; Karlin and Altschul 1990; Zeldovich et al. 2007). More than 100 hyperthermophilic or thermophilic bacterial and archaeal genomes have been sequenced and stored in public databases, thereby providing an unprecedented opportunity for studying the genetics, biochemistry, and evolution of these species, as well as for exploring the mechanisms of thermal adaptation. Many studies have been performed based on DNA sequence, protein sequence, certain protein families, and protein structure to investigate the mechanism of protein thermo-stability (Bae and Phillips 2004; Basak et al. 2007; Berezovsky and Shakhnovich 2005; Cambillau and Claverie 2000; De Vendittis et al. 2008). G+C content is not correlated significantly with optimal growth temperature (OGT). However, the increase of A+G content in coding genes can stabilize their thermo-stability, owing to the stacking effect of purines (Lao and Forsdyke 2000; Zeldovich et al. 2007). The thermo-stability of proteins is determined by a fine balance between many contributing factors, such as increment of hydrogen bonds, ion pairs, disulfide bridges or hydrophobic and aromatic interactions, changes in surface charge distribution, helix dipole stabilization, packing and reduction in solvent-accessible hydrophobic surface, contribution of specific chaperones, more compactness native conformation, variations of secondary structures, and so on (Basak and Ghosh 2005; De Vendittis and Bocchini 1996; Di Giulio 2000; Dong et al. 2008; Robb and Clark 1999; Tekaia and Yeramian 2006; Zeldovich et al. 2007). High-throughput comparative analysis of structures and complete genomes of several hyperthermophilic archaea has revealed that these organisms develop diverse strategies of thermophilic adaptation using two fundamental physical mechanisms (Berezovsky and Shakhnovich 2005). One is a 'structure-based' mechanism, i.e., some hyperthermostable proteins are significantly more compact than their mesophilic homologs, whereas no particular interaction type appears to cause stabilization (Bae and Phillips 2004; Berezovsky and Shakhnovich 2005; Dutta and Chaudhuri 2010). The other mechanism is a 'sequence-based' mechanism, i.e., hyperthermostable proteins do not show distinct structural differences from mesophilic homologs, whereas some apparently strong interactions, such as ionic interactions or additional salt bridges, are responsible for the high thermal stability of thermostable proteins (Berezovsky and Shakhnovich 2005; Makarova et al. 2003; Vetriani et al. 1998). Structure-stabilized proteins originate mostly from archaea, whereas sequence-stabilized proteins are mostly from bacteria (Berezovsky and Shakhnovich 2005).

Most previous studies have focused on certain protein families in different bacterial and archaeal genomes, ranging from psychrophiles to mesophiles to thermophiles/hyperthermophiles (Haney et al 1999; Georlette et al 2003; Bae and Phillips 2004). Although these studies show that the amino acid composition of many protein families is correlated with OGT, the proteomics characteristic of thermal adaptation remains unclear. The benefit of studying whole genomic sequences is that more fundamental properties of thermal adaptation can be obtained. Whole protein sequences of completely sequenced bacterial and archaeal genomes were compared instead of focusing on a single protein family. In total, more than 48,000 homologs from 526 bacterial and archaeal genomes were identified, and comprehensive comparison was carried out among the homologs distributed in four temperature range environments, i.e., hyperthermophilic (H, the organism grows above 75°C), thermophilic (T, the organism grows within the range 46°C to 75°C), mesophilic (M, the organism grows within the range 11°C to 45°C) and psychrophilic (P, the organism grows below 10°C). Analysis of these orthologs revealed a significant relationship between amino acid composition and thermal adaptation. The cluster of orthologous groups of proteins (COG) categories varied in the different temperature ranges (Tatusov et al. 2001). The current results strongly support the presence of a correlation between amino acid composition and thermostability, not only in a certain protein family, but also at the whole protein sequence level in prokaryotic genomes.

Materials and methods

Whole genomics sequences of 844 bacteria and archaea collected before the end of 2009 were downloaded from the NCBI Reference Sequence (RefSeq) project (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/). Among these, 526 genomes were retained by excluding the substrains in the same species. The classification, general features, and optimal growth temperatures of these species are listed in Supplemental Table 1. These genomes were classified into four temperature ranges according to the optimal growth temperature. Based on the information from the NCBI Entrez Genome Project (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi), 30, 51, 431, and 14 genomes were classified into hyperthermophiles (H, >75°C), thermophiles (T, 46°C to 75°C), mesophiles (M, 11°C to 45°C), and psychrophiles (P, <10°C).

Reciprocal-BLAST (E value $<10^{-10}$) was performed to obtain the orthologs of 526 bacterial and archaeal genomes (Altschul et al. 1990, 1994; Karlin and Altschul 1990). The distributions of each ortholog in the four temperature ranges were calculated. Three groups of datasets (HT-only group, M-only group, and Top-90 group) were picked out for further analysis. Orthologs containing less than ten protein sequences were excluded from the analysis. The HT-only group was defined as orthologs distributed either in thermophiles/hyperthermophiles that are not found in mesophiles and psychrophiles. The M-only group stands for orthologs distributed only in mesophiles and not found in thermophiles/hyperthermophiles and psychrophiles. The Top-90 group indicates orthologs distributed in at least 90% of the microbial genomes in the four temperature ranges. The program used to classify orthologs in the four temperature ranges was an in-house Perl script. The statistics of protein properties in orthologs were calculated with the PEPSTATS program in the EMBOSS package (Harrison 2000).

Results and discussion

Orthologs identified from 526 bacteria

General statistical results of the identified orthologs in the 526 bacterial and archaeal genomes are shown in Table 1. A total of 48,313 orthologs were identified, 18,042 of which contain more than ten proteins in their orthologous groups. Of these, 95 orthologs were found exclusively in the HT-only group, 1,659 in the M-only group, and 115 in the Top-90 group. The HT-only, M-only, and Top-90 group orthologs represent the proteins

Table 1 General statistics of orthologs in HT-only, M-only, and Top-90 groups

Orthologs	No.
Total	48,313
>10 proteins	18,042
HT-only	95
M-only	1,659
Top-90	115

Table 2 Functions and cluster of orthologous groups of proteins (COG) categories of HT-only orthologs

Function of HT-only orthologs	Number	COG
Information storage and processing	19	
30S ribosomal protein S27ae	1	COG1998J
50S ribosomal protein L13e	1	COG4352J
50S ribosomal protein L35Ae	1	COG2451J
50S ribosomal protein LX	1	COG2157J
Elongation factor 1-beta	1	COG2092J
Translation initiation factor, eIF-2B alpha subunit-related	1	COG1184J
Ribosomal biogenesis protein	1	COG2136JA
Hypothetical transcription regulator	1	COG1522K
ArsR family transcriptional regulator	1	COG1846K
Putative transcriptional regulator, CopG family	1	COG0864K
Small nuclear ribonucleoprotein (snRNP)-like protein	1	COG1958K
Transcriptional regulator, PadR-like family	1	COG1695K
Transcriptional regulators-like protein	1	COG1318K
TATA binding protein (TBP)-interacting protein (TIP49-like)	1	COG1224K
Endonuclease (RecB family)-like protein	1	COG4998L
CRISPR-associated autoregulator, DevR family	1	COG1857L
CRISPR -associated protein, family	1	COG1517L
DNA photolyase	1	COG1533L
DNA polymerase sliding clamp B1	1	COG0592L
Cellular processes and signaling	11	
S-layer domain protein precursor	1	COG1196D
Glycosyl transferase family protein	1	COG0463M
NAD-dependent epimerase/dehydratase	1	COG0451MG
Glycosyl transferase, family 39	1	COG4346O
AAA ATPase	1	COG0459O
Uncharacterized protein family UPF0033	1	COG0425O
UspA domain protein	1	COG0589T
MscS mechanosensitive ion channel	1	COG0668M
Small-conductance mechanosensitive channel-like protein	1	COG0668M
Membrane glycosyltransferase	1	COG1215M
Thioredoxin/glutaredoxin-like protein	1	COG5494O
Metabolism	8	
Membrane-anchored protein predicted to be involved in regulation of amylopullulanase	1	COG4945G
Dipeptidyl aminopeptidase/acylaminoacyl-peptidase-like protein	1	COG1506E
Amidophosphoribosyl transferase (ATASE)	1	COG0034F
Putative transcriptional regulator	1	COG0458EF
Hypothetical permease	1	COG0477GEPR
Major facilitator transporter	2	COG0477GEPR
Ferric uptake regulation protein	1	COG0735P
Poorly characterized	57	
Metallopeptidase-like protein	1	COG4900R
Fe-S oxidoreductase	1	COG5014R
HEPN domain-containing protein	1	COG1708R
Nucleotidyltransferase-like protein	1	COG4914R
RNA-binding protein	1	COG1818R
Tetratricopeptide TPR_2 repeat protein	1	COG0457R
Magnesium-dependent phosphatase-1	1	COG4996R
DNA polymerase beta subunit	4	COG1708R
AAA ATPase	1	COG0433R
Aminopeptidase Iap family-like protein	1	COG4882R
ATPase	1	COG1672R
Bacterio-opsin activator	1	COG3413R
P-loop ATPase/GTPase-like protein	1	COG4028R
PilT domain-containing protein	2	COG1848R
Putative nucleic acid-binding protein, contains PIN domain	1	COG4113R
tRNA m1G methyltransferase	1	COG2419S
paREP10	1	-
paREP6	1	-
ABC transporter ATP-binding protein, putative	1	COG4754S
CRISPR-associated protein, Csa1 family	3	COG4343S
HEPN domain-containing protein	2	COG2250S
Conserved crenarchaeal protein, putative	1	COG5493S
Thermopsin	1	-
PaREP1 family protein	4	-
Endo-1,4-beta-glucanase B	1	-
Hypothetical protein	22	-

unique in thermophiles/hyperthermophiles, mesophiles and in species that live in the four temperature ranges, respectively.

Functional analysis of the HT-only group and COG composition of the three ortholog groups

The functions of the 95 orthologs in the HT-only group are listed in Table 2. According to the COG functional category (http://www.ncbi.nlm.nih.gov/COG/old/palox.cgi?fun=all), 19, 11, 8, and 57 orthologs fall under 'information storage and processing', 'cellular processes and signalling', 'metabolism', and 'poorly characterized'. Sixty percent of orthologs in the HT-only group belong to the 'poorly characterized' category, which indicates that most of the HT-only orthologs have poor annotation information. The second largest category of HT-only orthologs is a group of 19 proteins classified under 'information storage and processing', of which 6 are ribosomal

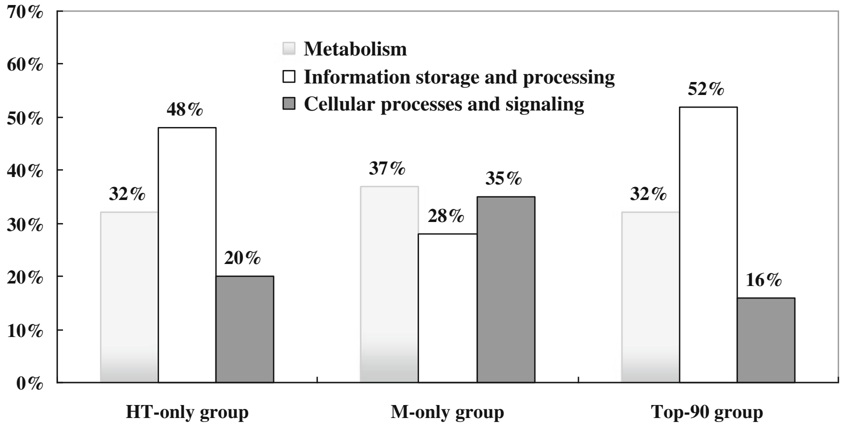

Fig. 1 Composition of cluster of orthologous groups of proteins (COG) of orthologs in HT-only, M-only, and Top-90 groups. The first, second, and third cluster of bars represent COG composition in HT-only, M-only, and Top-90 group, respectively. COG functional categories are as indicated in the figure

proteins and 7 are involved in transcriptional regulation. Although there are no particular functions ascribed to HT-only orthologs, their protein sequences are vastly different from those of mesophiles and psychrophiles. The HT-only orthologs play important roles in the thermal adaptation of thermophilic/hyperthermophilic species, which should be further investigated.

The COG functional categories in the three groups were compared. Except for the orthologs with no COG information or categorized under 'poorly characterized', the COG categories of orthologs in the HT-only group, M-only group and Top-90 groups were collected. The composition in each group is shown in Fig. 1. Bars in the first cluster represent the COG composition of the HT-only group, and the second and third clusters represent the composition of the M-only group and Top-90 group, respectively. Nearly half of the orthologs in the HT-only group and more than half of those in the Top-90 group were categorized under 'information storage and processing', whereas only 28% of those in the M-only group were classified as such. The proportion of orthologs classified under 'metabolism' was similar in the three groups, which accounted for about one-third of each ortholog group. The proportion falling under 'cellular processes and signalling' varies in the three groups. Orthologs in the M-only group have significantly higher proportion (35%), followed by the HT-only group (20%) and Top-90 group (16%). It has been hypothesized that in order to maintain the survival of organisms in high-temperature environments, thermophiles/hyperthermophiles evolved more functional proteins for 'information storage and processing'.

Amino acid composition and protein property analysis

Amino acid composition in the three groups was calculated, and the corresponding comparison between the HT-only and M-only groups is listed in Table 3. Significant differences were found. More charged (D+E+H+K+R, 24.95% vs. 20.25%; $P<0.001$), acidic (D+E, 9.67% vs. 7.48%; $P<0.001$) and basic (H+K+R, 15.28% vs. 12.77%; $P<0.001$) amino acid residues were found in the HT-only orthologs. Other research groups have also observed that thermostable proteins have an amino acid composition biased for enhancing electrostatic or hydrophobic interactions (De Farias and Bonato 2002; De

Table 3 Comparison of protein properties between HT-only and M-only groups

Protein property[a]	HT-only group	M-only group	*P*-value
No. of amino acids	235.66 (±133.44)	273.58 (±195.00)	0.010
Average residue weight	113.38 (±2.58)	111.14 (±3.00)	0
Isoelectric point	7.58 (±1.76)	7.17 (±1.75)	0.026
Tiny amino acids (%)	24.51 (±3.57)	29.44 (±4.37)	0
Small amino acids (%)	43.96 (±4.51)	48.87 (±5.03)	0
Aliphatic amino acids (%)	38.64 (±4.07)	37.95 (±4.89)	0.075
Aromatic amino acids (%)	9.79 (±2.82)	9.88 (±3.01)	0.781
Non-polar amino acids (%)	40.78 (±4.95)	41.05 (±5.28)	0.658
Polar amino acids (%)	52.71 (±5.96)	52.27 (±6.01)	0.609
Charged amino acids (%)	24.95 (±6.54)	20.25 (±5.46)	0
Basic amino acids (%)	15.28 (±4.08)	12.77 (±3.28)	0
Acidic amino acids (%)	9.67 (±3.52)	7.48 (±3.31)	0

[a] The amino acids groups are classified as follows: tiny amino acids (A+C+G+S+T), small amino acids (A+C+D+G+N+P+S+T+V), aliphatic amino acids (A+I+L+V+G), aromatic amino acids (F+W+Y), non-polar amino acids (A+F+I+L+M+P+V+W), polar amino acids (C+G+Y+S+T+N+Q+D+E+H+K+R), charged amino acids (D+E+H+K+R), basic amino acids (H+K+R), acidic amino acids (D+E)

Venditis et al. 2008; Di Giulio 2000; Farias and Bonato 2003; Zeldovich et al. 2007). The M-only set has significantly higher proportions of tiny (A+C+G+S+T, 29.44% vs 24.51%; $P<0.001$) and small (A+C+D+G+N+P+S+T+V, 48.87% vs 43.96%; $P<0.001$) amino acid residues. Differences found in proportions of aliphatic, aromatic, non-polar, and polar amino acids are not significant. The average length of HT-only orthologs tends to be shorter than those of M-only orthologs, with an average length of 236 vs 274 amino acids, respectively ($P<0.05$). The protein properties can be attributed to the maintenance of the thermo-stability of thermophilic/hyperthermophilic genomes in high-temperature environments.

Many protein properties in the Top-90 group of orthologs in the four temperature ranges were also calculated and compared (Fig. 2a–h). The protein properties investigated include the number of amino acid residues (Fig. 2a), average residue

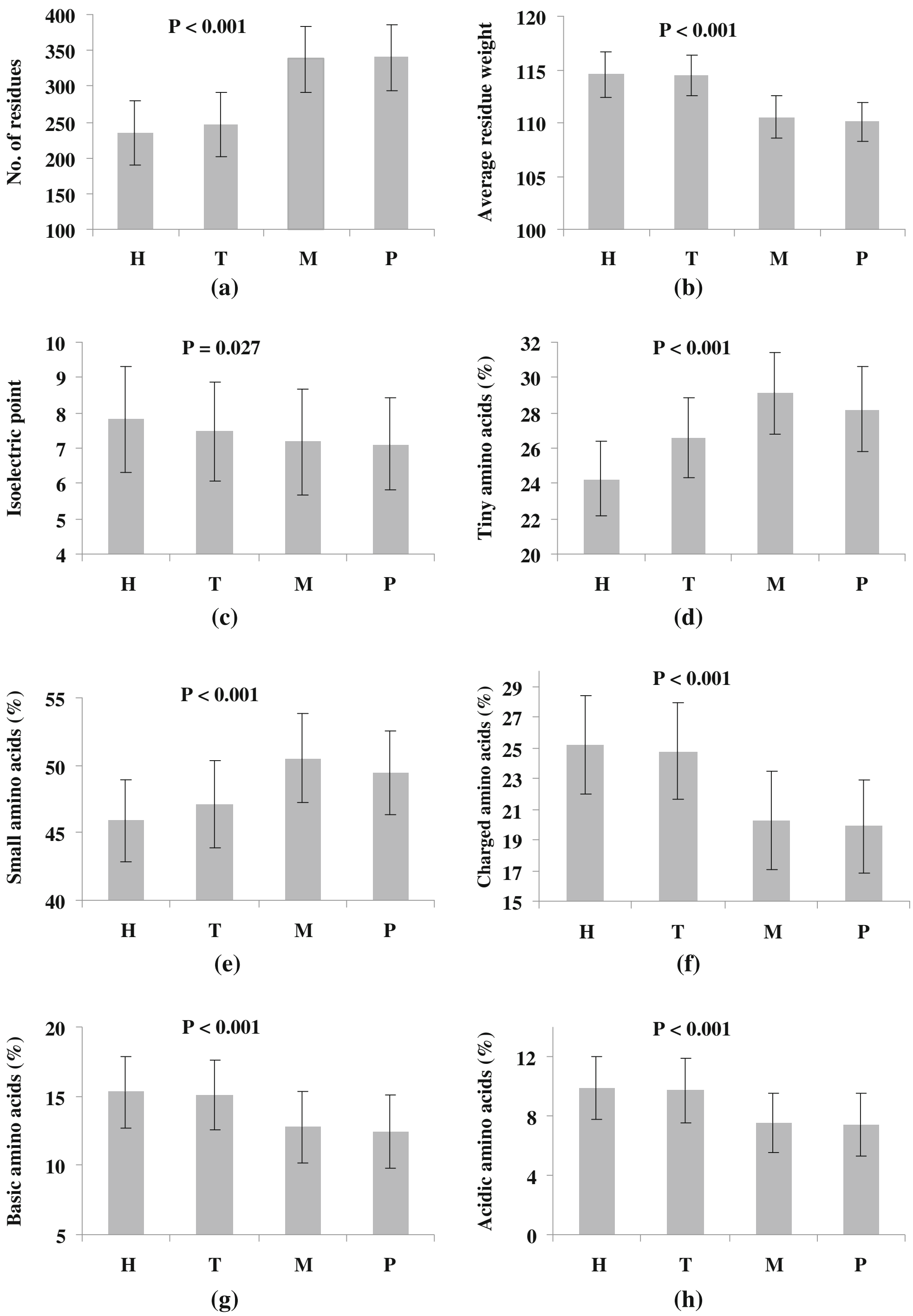

Fig. 2 Comparison of the protein properties among orthologs in the Top-90 group in hyperthermophiles, thermophiles, mesophiles, and psychrophiles. The calculated protein properties include number of **a** residues, **b** average residue weight, **c** isoelectric point, **d** tiny amino acids, **e** small amino acids, **f** charged amino acids, **g** basic amino acids, and **h** acidic amino acids. *H* Hyperthermophiles, *T* thermophiles, *M* mesophiles, *P* psychrophiles

weight (Fig. 2b), isoelectric point (Fig. 2c), tiny amino acids (A+C+G+S+T, Fig. 2d), small amino acids (A+C+D+G+N+P+S+T+V, Fig. 2e), charged amino acids (D+E+H+K+R, Fig. 2f), basic amino acids (H+K+R, Fig. 2g), and acidic amino acids (D+E, Fig. 2h). In each histogram, the columns indicate average value and the error bars indicates standard deviation. Most of the above mentioned protein properties differ significantly among hyperthermophiles, thermophiles, and mesophiles, except isoelectric point. However, the differences between mesophiles and psychrophiles are not significant. In the protein properties in the HT-only group, there are significantly higher proportions of charged, basic, and acidic amino acids in hyperthermophilic and thermophilic genomes than in mesophilic and psychrophilic genomes, as well as lower proportions of tiny and small amino acids. The same trends are also observed between the HT-only group and M-only group of orthologs. The average protein sequence length of the orthologs also differs among the different temperature ranges, and the bacterial and archaeal genomes of organisms that live in higher-temperature environments tend to have the shortest average protein sequence lengths.

Conclusion

The current research confirmed the previous findings of a significant relationship between amino acid composition and protein thermostability at the level of whole prokaryotes protein sequences. The proteins from thermophiles/hyperthermophiles tend to use more charged (D+E+H+K+R) amino acids to adapt to high temperature environments. The average length of protein sequences in thermophiles/hyperthermophiles is much shorter than those in mesophiles. The differences of amino acid composition between thermophiles/hyperthermophiles and mesophiles might be beneficial for the maintenance and stability of proteins in high temperatures. Furthermore, the differences of COG composition showed that more functional proteins participate in ‘information storage and processing’ in thermophiles/hyperthermophiles, thus enabling the organisms to adapt to the high-temperature environments.

Acknowledgments The present study was supported by the National Natural Science Foundation of China (31071659), the National High Technology Research and Development Program of China (2008AA09Z411), and the Fundamental Research Funds for the Central Universities (2011PY071).

References

Altschul SF, Boguski MS, Gish W, Wootton JC (1994) Issues in searching molecular sequence databases. Nat Genet 6:119–129

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. Mol Biol 215:403–410

Bae E, Phillips GN Jr (2004) Structures and analysis of highly homologous psychrophilic, mesophilic, and thermophilic adenylate kinases. J Biol Chem 279:28202–28208

Basak S, Ghosh TC (2005) On the origin of genomic adaptation at high temperature for prokaryotic organisms. Biochem Biophys Res Commun 330:629–632

Basak S, Roy S, Ghosh TC (2007) On the origin of synonymous codon usage divergence between thermophilic and mesophilic prokaryotes. FEBS Lett 581:5825–5830

Berezovsky IN, Shakhnovich EI (2005) Physics and evolution of thermophilic adaptation. Proc Natl Acad Sci USA 102:12742–12747

Burra PV, Kalmar L, Tompa P (2010) Reduction in structural disorder and functional complexity in the thermal adaptation of prokaryotes. PLoS One 5:e12069

Cambillau C, Claverie JM (2000) Structural and genomic correlates of hyperthermostability. J Biol Chem 275:32383–32386

De Farias ST, Bonato MC (2002) Preferred codons and amino acid couples in hyperthermophiles. Genome Biol 3:PREPRINT0006

Dehouck Y, Folch B, Rooman M (2008) Revisiting the correlation between proteins' thermoresistance and organisms' thermophilicity. Protein Eng Des Sel 21:275–278

De Vendittis E, Bocchini V (1996) Protein-encoding genes in the sulfothermophilic archaea Sulfolobus and Pyrococcus. Gene 176:27–33

De Vendittis E, Castellano I, Cotugno R, Ruocco MR, Raimo G, Masullo M (2008) Adaptation of model proteins from cold to hot environments involves continuous and small adjustments of average parameters related to amino acid composition. J Theor Biol 250:156–171

Di Giulio M (2000) The late stage of genetic code structuring took place at a high temperature. Gene 261:189–195

Dong H, Mukaiyama A, Tadokoro T, Koga Y, Takano K, Kanaya S (2008) Hydrophobic effect on the stability and folding of a hyperthermophilic protein. J Mol Biol 378:264–272

Dutta A, Chaudhuri K (2010) Analysis of tRNA composition and folding in psychrophilic, mesophilic and thermophilic genomes: indications for thermal adaptation. FEMS Microbiol Lett 305:100–108

Farias ST, Bonato MC (2003) Preferred amino acids and thermostability. Genet Mol Res 2:383–393

Haney PJ, Badger JH, Buldak GL, Reich CI, Woese CR, Olsen GJ (1999) Thermal adaptation analyzed by comparison of protein sequences from mesophilic and extremely thermophilic Methanococcus species. Proc Natl Acad Sci USA 96:3578–3583

Harrison RG (2000) Expression of soluble heterologous proteins via fusion with NusA protein. Innovations 11:4–7

Georlette D, Damien B, Blaise V, Depiereux E, Uversky VN, Gerday C, Feller G (2003) Structural and functional adaptations to extreme temperatures in psychrophilic, mesophilic, and thermophilic DNA ligases. J Biol Chem 278:37015–37023

Karlin S, Altschul SF (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc Natl Acad Sci USA 87:2264–2268

Lao PJ, Forsdyke DR (2000) Thermophilic bacteria strictly obey Szybalski's transcription direction rule and politely purine-load RNAs with both adenine and guanine. Genome Res 10:228–236

Makarova KS, Wolf YI, Koonin EV (2003) Potential genomic determinants of hyperthermophily. Trends Genet 19:172–176

Robb FT, Clark DS (1999) Adaptation of proteins from hyperthermophiles to high pressure and high temperature. J Mol Microbiol Biotechnol 1:101–105

Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res 29:22–28

Tekaia F, Yeramian E (2006) Evolution of proteomes: fundamental signatures and global trends in amino acid compositions. BMC Genomics 7:307

Vetriani C, Maeder DL, Tolliday N et al (1998) Protein thermostability above 100°C: a key role for ionic interactions. Proc Natl Acad Sci USA 95:12300–12305

Zeldovich KB, Berezovsky IN, Shakhnovich EI (2007) Protein and DNA sequence determinants of thermophilic adaptation. PLoS Comput Biol 3:e5