Amino Acid Coupling Patterns in Thermophilic Proteins

Han-Kuen Liang,[1,2†] Chia-Mao Huang,[2,3†] Ming-Tat Ko,[2*] and Jenn-Kang Hwang[1*]
[1]*Institute of Bioinformatics, National Chiao Tung University, HsinChu, Taiwan*
[2]*Institute of Information Science, Academia Sinica, NanKang, Taipei, Taiwan*
[3]*Department of Computer Science, National Tsing Hua University, HsinChu, Taiwan*

ABSTRACT **Structural analysis is useful in elucidating structural features responsible for enhanced thermal stability of proteins. However, due to the rapid increase of sequenced genomic data, there are far more protein sequences than the corresponding three-dimensional (3D) structures. The usual sequence-based amino acid composition analysis provides useful but simplified clues about the amino acid types related to thermal stability of proteins. In this work, we developed a statistical approach to identify the significant amino acid coupling sequence patterns in thermophilic proteins. The amino acid coupling sequence pattern is defined as any 2 types of amino acids separated by 1 or more amino acids. Using this approach, we construct the ρ profiles for the coupling patterns. The ρ value gives a measure of the relative occurrence of a coupling pattern in thermophiles compared with mesophiles. We found that thermophiles and mesophiles exhibit significant bias in their amino acid coupling patterns. We showed that such bias is mainly due to temperature adaptation instead of species or GC content variations. Though no single outstanding coupling pattern can adequately account for protein thermostability, we can use a group of amino acid coupling patterns having strong statistical significance (*p* values < 10^{-7}) to distinguish between thermophilic and mesophilic proteins. We found a good correlation between the optimal growth temperatures of the genomes and the occurrences of the coupling patterns (the correlation coefficient is 0.89). Furthermore, we can separate the thermophilic proteins from their mesophilic orthologs using the amino acid coupling patterns. These results may be useful in the study of the enhanced stability of proteins from thermophiles—especially when structural information is scarce. Proteins 2005;59:58–63.**

Key words: amino acid coupling patterns; thermophilic proteins; mesophilic proteins; thermal stability

INTRODUCTION

Structural analysis[1–24] has identified a number of structural features responsible for thermal stability of proteins; these structural features include tighter hydrophobic packing[3,4,6,18]; more charged amino acids or salt bridges,[3,4,13,18] or more charged residues at the protein surface[8,19]; the preferential arrangement of charged residues with a 1-4 helical spacing[9,11,19]; the stabilization of the helices by an increase in negative charge at the N-terminal and an increase in helical content[19]; a shortened loop[4]; a decrease in the entropy of unfolding[18]; less free cysteine amino acids except those involved in disulfide bridges and metal binding, or those inaccessible to the solvent[18,20]; higher proportions of charged versus polar (noncharged) amino acids,[5,8] or asymmetrical substitution patterns for certain amino acid pairs[25]; and more side-chain–side-chain hydrogen bonds.[6,11] However, due to the advances of genomic research in recent years, the accumulation of protein sequences far outpaces that of the corresponding three-dimensional (3D) structures. Hence, sequence-based analysis is still valuable in the study of thermal stability of proteins. The often-used sequence-based methods[5,7,14,26–28] differentiate the amino acid compositions between thermophilic and mesophilic proteins, and show that thermophilic proteomes exhibit significant bias in their amino acid compositions. For example, Val, Glu, and total charged residue content are found to be higher, while polar amino acids like Gln, Asn, Ser, Thr, and His contents are significantly lower in thermophilic genomes. However, amino acid composition analysis provides a useful but simplified picture of the relative importance of each individual amino acid type in the thermophilic proteins. Such analysis overlooks the coupling effects between amino acid types on thermal stability of proteins. In this article, we study the amino acid coupling sequence patterns for a data set comprising 74 mesophilic and 15 thermophilic genomes. We developed a statistical approach to analyze the amino acid coupling patterns in thermophilic proteins. We also discuss the structural implications of these coupling

The Supplementary Materials referred to in this article can be found at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/index.html

Grant sponsor: National Science Council, Taiwan (to J.-K. Hwang and M.-T. Ko).

†Han-Kuen Liang and Chia-Mao Huang contributed equally to this article.

*Correspondence to: Jenn-Kang Hwang, Institute of Bioinformatics, National Chiao Tung University, HsinChu 30015, Taiwan. E-mail: jkhwang@cc.nctu.edu.tw. Ming-Tat Ko, Institute of Information Science, Academia Sinica, NanKang, Taipei 115, Taiwan. E-mail: mtko@iis.sinica.edu.tw

Received 10 May 2004; Accepted 15 October 2004

Published online 1 February 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20386

patterns and their roles in thermal stabilization of proteins.

MATERIAL AND METHODS

Let [*XdZ*] denote the amino acid coupling pattern of amino acid types *X* and *Z* that are separated by *d* amino acids. Since the protein sequence is directional, the sign of *d* is determined by the relative positions of *X* and *Z*. If *X* is closer to the N-terminal side, *d* is defined as positive, and if *X* is closer to the C-terminal side, it is defined as negative. Let $N(XdZ)$ be the number of occurrences of the pattern [*XdZ*]. We define the conditional probability R_{XdZ} as

$$R_{XdZ} = \frac{N(XdZ)}{N(Xd\cdot)}, \tag{1}$$

where $N(Xd\cdot) = \sum_Y N(XdY)$ and $Y \in \{20 \text{ types of amino acid}\}$. R_{XdZ} is the probability of amino acid *Z* occurring at *d* amino acids from amino acid *X*. The coupling strength C_{XdZ} between *X* and *Z* of the pattern [*XdZ*] is given by

$$C_{XdZ} = \frac{R_{XdZ}}{P(Z)}, \tag{2}$$

where $P(Z)$ is the probability of the occurrence of amino acid *Z*. C_{XdZ} indicates the coupling strength of amino acid *Z* at *d* amino acids from amino acid *X*. If $C_{XdZ} \geq 1$, then *X* and *Z* are positively correlated with respect to the distance *d*, and if $C_{XdZ} < 1$, they are negatively correlated. We use $\bar{R}^T_{XdZ}$ and $\bar{R}^M_{XdZ}$ to denote the means of R_{XdZ} over thermophilic and mesophilic proteins, respectively. Similarly, $\bar{C}_{XdZ}$ denotes the mean of C_{XdZ} over all proteins. To compute the relative occurrence of [*XdZ*] in thermophilic proteins, we define

$$\rho_{XdZ} = \frac{\bar{R}^T_{XdZ}}{\bar{R}^M_{XdZ}}. \tag{3}$$

The ρ value of pattern [*XdZ*] gives a measure of its relative occurrence in thermophiles compared with mesophiles. If $\rho_{XdZ} > 1$, [*XdZ*] is increased in thermophilic proteins, and if $\rho_{XdZ} < 1$, it is decreased in thermophilic proteins. We will refer to ρ_{XdZ} as the thermophilic coefficient, or simply the ρ value of [*XdZ*]. To check the statistical significance of [*XdZ*], we carry out a statistical test on R_{XdZ} and C_{XdZ} between thermophilic and mesophilic genomes. The resultant *p* value is used to determine whether the null hypothesis is true. For example, in the statistical test, if the *p* value is less than 10^{-2}, we have 99% confidence that the coupling patterns present in the thermophilic and mesophilic samples are significantly different. Since the sample size of the amino acid coupling pattern [*XdZ*] is too small for a confident normal distribution test, we carry out the nonparametrical Wilcoxon rank-sum test on R_{XdZ} and C_{XdZ} between thermophilic and mesophilic genomes to check the statistical significance of [*XdZ*]. The advantage of nonparametric methods (such as the Wilcoxon rank-sum test or the equivalent Mann–Whitney test) over their parametric counterparts (such as the *t* test) is the absence of assumptions regarding the sample distribution. A nonparametric test is more powerful with small sample sizes and with non-normal data.[29] In fact, the Wilcoxon procedure is applicable to both small and large samples, and its advantage actually increases when the sample size becomes larger. For samples having the normal distribution, the nonparametric test will have less power and is less likely to give small *p* values, especially when the sample size is small; however, this is not the case for the present study (see Results and Discussion section). We have studied $20 \times 20 \times 40 = 16{,}000$ amino acid-coupling patterns [*XdZ*] for *X*, *Z* over all 20 types of amino acid and $-20 \leq d \leq 20$. When the separation is greater than 20 amino acids, we find that $C_{XdZ} \sim 1$, indicating that the correlation between amino acids becomes insignificant when $|d| \geq 20$. The *p* values of the Wilcoxon rank-sum test for R_{XdZ} and C_{XdZ} are called their $RS(R)$ and $RS(C)$ values, respectively. Eqs. (1) and (2) are formulated for a single-residue coupling sequence pattern. It is not hard to generalize these equations for a group of residue coupling sequence patterns denoted by Ω. The generalized equations are given by

$$R_\Omega = \frac{\sum_{XdY\in\Omega} Z(R_{XdY}) \times \zeta(\bar{R}^T_{XdY} - \bar{R}^M_{XdY})}{|\Omega|}, \tag{4}$$

$$C_\Omega = \frac{\sum_{XdY\in\Omega} Z(C_{XdY}) \times \zeta(\bar{C}^T_{XdY} - \bar{C}^M_{XdY})}{|\Omega|}, \tag{5}$$

where $|\Omega|$ is the number of the sequence patterns of the set Ω, and $Z(R_{XdY})$ and $Z(C_{XdY})$ are the standardized normal scores based on standard normal distribution $N(0,1)$, and the function $\zeta(x)$ gives -1, 0, or 1, depending on whether *x* is negative, zero, or positive.

DATA SET

Our data set comprises 89 prokaryotic genomes, 15 archaea, and 74 bacteria, obtained from the Comprehensive Microbial Resource (CMR) of TIGR database (http://www.tigr.org). Using optimal growth temperature (OGT) 45°C as the thermophilic delimitation, we further divide this data set into two parts: the thermophilic set containing 15 genomes, which comprises 12 archeae and 3 bacteria, and the mesophilic set containing 74 genomes, which comprises 3 archeae and 71 bacteria. The OGTs of the 89 prokaryotes are available at http://pgtdb.csie.ncu.edu.tw.

RESULTS AND DISCUSSION

ρ Profiles of Amino Acid Coupling Patterns

Using Eq. (3), we are able to construct the ρ profile of the amino acid coupling patterns. The ρ profile is useful in providing a global picture of the relative occurrences of the coupling pattern in thermophiles compared with mesophiles. An example of the ρ profile for [*xdC*] is shown in Figure 1(A), which shows the ρ values, together with $RS(R)$ and $RS(C)$. Most [*xdC*]s have $\rho < 1$ and, hence, appear to be decreased in thermophiles. These results are consistent with previous reports[14,18,19] that the Cys composition is in general decreased in thermophiles. However, we note that there exist some statistically significant

patterns with $\rho > 1.4$ [indicated by the arrow in Fig. 1(A)]. We zoom in on this region in Figure 1(B). These patterns are mostly of the form [*CdC*], some instances of which are [*C*3*C*], [*C*4*C*], and [*C*7*C*]. Rosato et al.[20] previously reported that cysteine clustering is closely related to the growth temperature of the organism. Structural analysis[20] showed that the increased stability of the cysteine clusters is probably due to their involvement in coordination of metal ions such as zinc, iron, or FeS groups, or in disulfide bonds. This example shows that our approach is able to identify and provide a more detailed description of sequence features in thermophilic proteins than the conventional composition analysis. In the following sections, we discuss the ρ profiles of coupling patterns of the general coupling pattern [*xdZ*], where *Z* denotes the particular amino acid type and *x* is any amino acid type.

ρ Profiles of [*xdZ*]

The figures of the ρ profiles of the coupling patterns of the form [*xdZ*] can be found in the Supplementary Material.

[*xdE*] and [*xdV*]

We found that [*xdE*] occurs more in thermophiles, and this observation is consistent with previous reports[14,19] that Glu content is usually higher in the thermophilic proteins. Specifically, the most statistically significant [$RS(C)$ and $RS(R) < 10^{-5}$] patterns are [*KdE*], [*RdE*], [*EdE*], and [*DdE*]. The first 3 patterns usually occur in helices when $d = 3$. For example, using the nonredundant Protein Data Bank set (nr-PDB) with sequence homologs removed using the sequence-similarity cutoff BLAST *p* value $= 10^{-7}$, we found that 58% of both [*K*3*E*] and [*R*3*E*], and 53% of [*E*3*E*] occur in helices. These results are in accordance with the previous report[19] that both local salt bridges and helical conformations are significantly increased in thermophilic proteins.

The ρ profile of [*xdV*] (see Supplementary Material) is similar to that of [*xdE*], though the nonpolar valine and the charged glutamate are completely different types of amino acids. These are the coupling patterns [*DdV*], [*KdV*], [*NdV*], and [*YdV*] that are significantly increased in thermophiles. The structural implications of these patterns are not clear, though these patterns frequently occur in α-helices or β-sheets, and a higher proportion of secondary structures is known to be an important contributor to increased thermal stability.[19]

[*xdP*] and [*xdC*]

The ρ profile of [*xdP*] is similar to that of [*xdC*]. Most instances of [*xdP*] are increased in thermophiles ($\rho > 1$), though with relatively high *p* values. It is reported[19] that the Pro composition is increased in thermophilic proteins. There exist a few statistical significant patterns with $\rho > 1.4$ (indicated by the arrow in the figure in the Supplementary Material), which are [*CdP*] and [*PdP*]. We found from structural analysis that [*PdP*]s (or proline clusters) are often involved in the formation of the polyproline II helix. The helical conformation, together with the reduced conformational entropy, may contribute to protein stability.

[*xdQ*], [*xdT*], and [*xdH*]

The coupling patterns involving polar amino acids are usually decreased in thermophiles. It is reported[5,14,18] that the Gln composition, as well as other polar amino acids like Ser, Gln, Asn, Thr, and Cys, are decreased in thermophiles. Specifically, the coupling patterns with *p* values $< 10^{-6}$ are [*EdQ*], [*GdQ*], [*RdQ*], and [*QdQ*]. The homo-amino acid coupling pair [*QdQ*] presents a special case in sequence coupling patterns. Figure 2 shows the homo-amino acid coupling patterns for 20 amino acid types. Only [*EdE*], [*CdC*], and [*PdP*] show statistically significant instances that are increased in thermophilic proteins (see also the previous sections).

Most instances of [*xdT*] have $\rho < 1$. We notice that the particular pattern [(charged residue)*dT*] is significantly decreased in thermophiles. For example, [*E*3*T*] has $\rho = 0.72$, $RC(R) = 9.6 \times 10^{-8}$ and $RC(C) = 2.4 \times 10^{-8}$. Though Glu is usually increased in thermophiles, the coupling pattern [*E*3*T*] is in fact decreased in thermophiles. The ρ profile of [*xdH*] shows a similar shape to that of [*xdT*]. Interestingly, we observe that [(charged residue)*dH*] is also significantly decreased in thermophilic proteins.

Other Coupling Patterns

[*xdL*] does not show any significant bias toward thermophiles. However, a particular instance, [*CdL*], is decreased in thermophiles with statistical significance. Other patterns like [*xdM*], [*xdF*], [*xdW*], and [*xdG*] also show similar neutral ρ profiles. [*xdI*], unlike [*xdL*], is increased in thermophiles. For the patterns involving aromatic amino acids, [*xdF*] and [*xdW*] are decreased in thermophilic proteins, but [*xdY*] is increased. For patterns involving charged amino acids, [*xdE*], [*xdK*], and [*xdR*] are increased in thermophilic proteins, but interestingly, [*xdD*] is decreased. For patterns involving polar amino acids, [*xdS*] and [*xdN*] are in general decreased in thermophilic proteins. The ρ profile of [*xdA*] pattern is similar to that of [*xdN*] and is decreased in thermophiles, despite the fact that alanine and asparagine are 2 different types of amino acids.

Significant Amino Acid Coupling Patterns

The net thermal stability of proteins usually results from a multitude of different coupling patterns, and no single outstanding sequence or structural feature can adequately account for thermophilic proteins. We identify from the amino acid coupling pattern the most significant ones with *p* values $< 10^{-7}$ for both $RS(R)$ and $RS(C)$. We denote this set by Ω, which contains the following thermophilic amino acid coupling patterns: [*C*(−2)*P*], [*C*1*P*], [*C*3*C*], [*C*4*C*], [*C*6*C*], [*C*7*C*], [*K*(−7)*E*], [*K*(−4)*E*], [*K*3*E*], [*K*4*E*], and [*H*(−4)*V*], and the following mesophilic amino acid coupling patterns: [*C*(−4)*L*], [*C*(−3)*L*], [*C*(−2)*L*], [*C*2*L*], [*C*3*L*], [*D*(−5)*T*], [*D*(−4)*T*], [*E*(−8)*T*], [*E*(−4)*T*], [*E*1*Q*], [*E*3*T*], [*E*4*T*], [*G*(−3)*Q*], [*K*(−4)*T*], [*K*2*T*], and [*K*3*T*].

Identification of Thermophilic and Mesophilic Proteins Using Amino Acid Coupling Patterns

Most sequenced thermophilic genomes are archaea (as also reflected in our thermophile data set—12 archaea and 3 bacteria), and it is possible that some of the amino acid coupling patterns between thermophilic and mesophilic proteins may be due to phylogenetic differences instead of temperature adaptation. We compute C_Ω for the set Ω using Eq. (5) for both bacteria and archaea. Figure 3 shows the C_Ω–OGT plot for both archaea and bacteria genomes. The amino acid coupling patterns can clearly distinguish between thermophiles and mesophiles of both bacteria and archaea. The results show that we can identify the amino acid coupling patterns that are indeed due to temperature adaptation. Furthermore, we observe a good linear correlation between C_Ω and OGT (the correlation coefficient is 0.89). This is encouraging, since the linear relationship is obtained without adjustable parameters.[28]

To distinguish thermophilic proteins and their mesophilic orthologs presents a much harder challenge, because these orthologs usually share higher degrees of sequence similarity. Define τ and μ as the occurrences of thermophilic and mesophilic amino acid patterns of the set Ω, respectively. We compute τ and μ for both thermophilic and mesophilic orthologs of the clusters of orthologous groups (COG) families.[30] To avoid sampling bias toward either thermophiles or mesophiles, we selected the COG groups under the consideration that each COG group should cover both bacteria and archaea and have proper proportions of thermophiles and mesophiles. Since our sequence patterns are sampled from prokaryotes, we exclude the eukaryotic sequences, if any, from the COG groups. The resultant selected GOG groups are COG0003, COG0068, COG0121, COG0156, COG0430 and COG1042. The τ–μ plot of these COG families is shown in Figure 4, with each point (τ,μ) representing 1 ortholog. Thermophilic proteins are generally well separated from their mesophilic orthologs, with most thermophilic orthologs clustering in the lower right area of the τ–μ region, while the mesophilic orthologs cluster in the upper left regions.

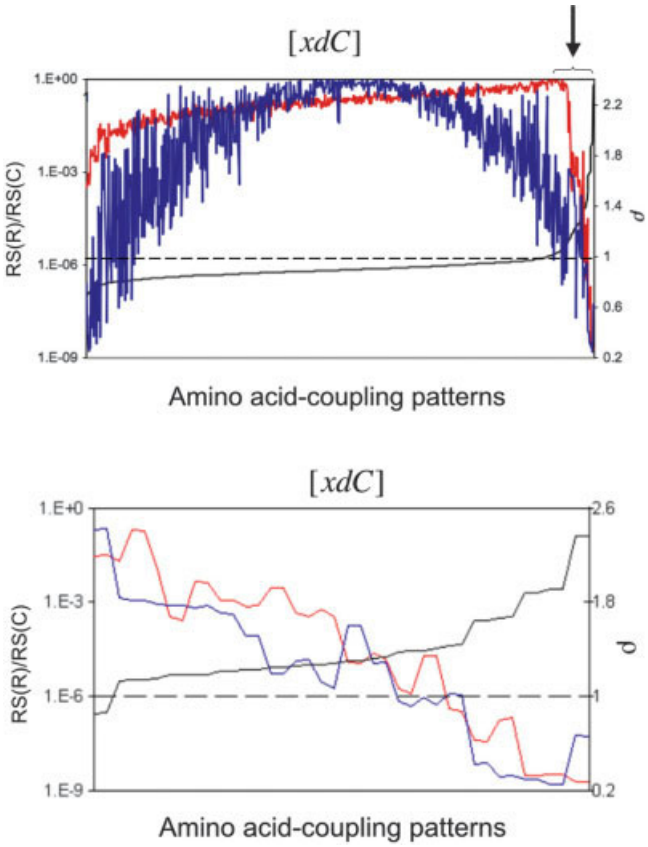

Fig 1. (**A**) Theρ, *RS*(*R*), and *RS*(*C*) profiles of the amino acid coupling pattern [*xdC*]. The ρ values are plotted in black (scale on the right), and the *RS*(*R*), and *RS*(*C*) values in red and blue, respectively (logarithmic scale on the left). The abscissas are the amino acid coupling patterns [*xdC*] sorted according to ascending ρ values. The dotted line indicates the threshold $\rho = 1$. The arrow indicates the region of the statistical significant patterns with $\rho > 1.4$. (**B**) The zoom-in view of this region.

GC Content and Amino Acid Coupling Patterns

Though GC content is the dominant influence on amino acid composition, it has been shown that GC pressure and

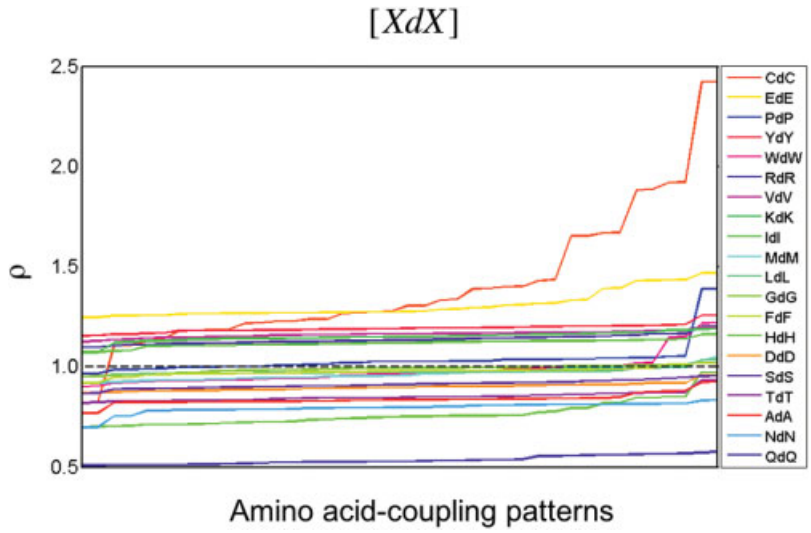

Fig. 2. The ρ profiles of 20 homo-amino acid coupling patterns [*XdX*].

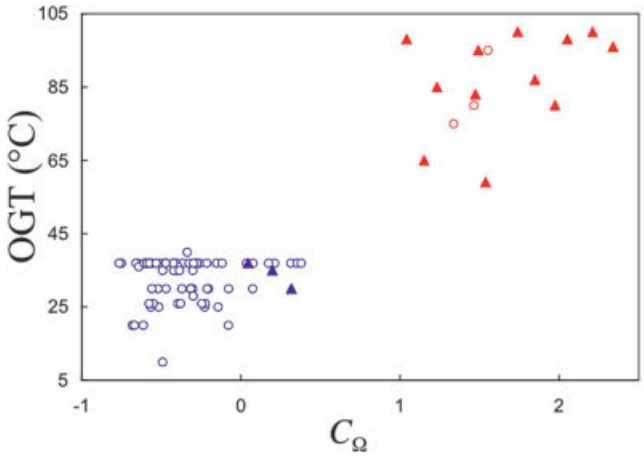

Fig. 3. The C_Ω–OGT plot of the thermophiles and mesophiles. The circles represent bacterial genomes and the triangle the archaea genomes. Thermophiles are colored in red, and mesophiles in blue.

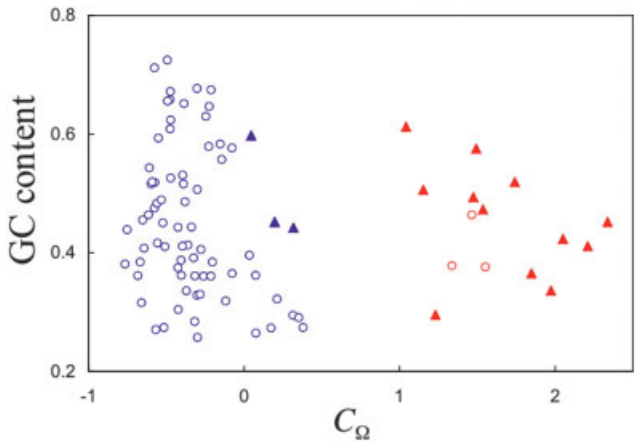

Fig. 5. The C_Ω–GC content plot of the thermophiles and mesophiles. The circles represent bacterial genomes and the triangles, the archaea genomes. Thermophiles are colored in red, and mesophiles in blue.

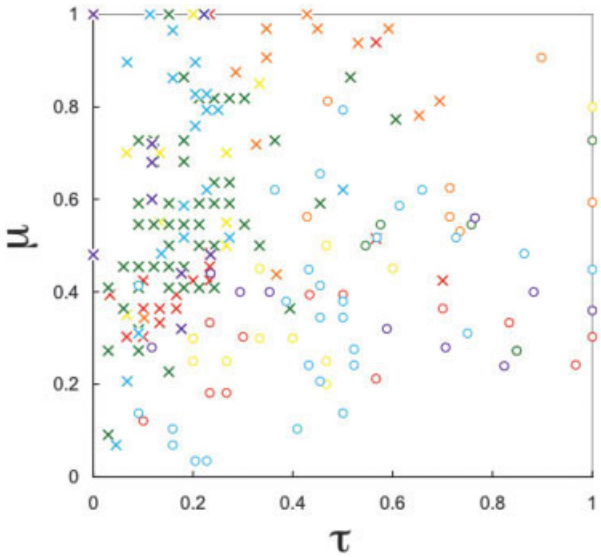

Fig. 4. The τ–μ plot of the COG families: COG0003 (red), COG0068 (orange), COG0121 (yellow), COG0156 (green), COG0430 (violet) and COG1042 (turquoise). The thermophilic proteins are shown in circles and the mesophilic orthologs in crosses. The occurrences of the thermophilic and mesophilic amino acid coupling patterns are normalized by dividing the maximal occurrences of the corresponding patterns of each COG family.

thermophily are essentially independent of each other.[14] We compute C_Ω for both bacteria and archaea. Figure 5 compares C_Ωs and the corresponding GC contents of the genomes. While C_Ω clearly distinguishes between thermophiles and mesophiles, both thermophiles and mesophiles scatter over a range of similar the GC contents.

CONCLUSIONS

We present a statistical analysis of the relationship between coupling patterns and thermophily of genomes. Though no single outstanding pattern can adequately account for protein thermophily, it is possible to distinguish between thermophiles and mesophiles using a set of thermophilic and mesophilic amino acid coupling patterns. However, despite the dominant influence of GC content on amino acid composition, we found that that GC content and thermophily are essentially independent of each other, as previously reported.[14] Furthermore, we are able to distinguish thermophilic proteins from their mesophilic orthologs using these amino acid coupling patterns. Our approach may be useful in elucidating the relationship between sequence features and protein thermal stability.

ACKNOWLEDGMENT

We are grateful for both hardware and software support of the Structural Bioinformatics Core in Hsin Chu.

REFERENCES

1. Tanaka A, Flanagan J, Sturtevant JM. Thermal unfolding of staphylococcal nuclease and several mutant forms thereof studied by differential scanning calorimetry. Protein Sci 1993;2:567–576.
2. Delboni LF, Mande SC, Rentier-Delrue F, Mainfroid V, Turley S, Vellieux FM, Martial JA, Hol WG. Crystal structure of recombinant triosephosphate isomerase from *Bacillus stearothermophilus*: an analysis of potential thermostability factors in six isomerases with known three-dimensional structures points to the importance of hydrophobic interactions. Protein Sci 1995;4:2594–2604.
3. Lim JH, Yu YG, Han YS, Cho S, Ahn BY, Kim SH, Cho Y. The crystal structure of an Fe-superoxide dismutase from the hyperthermophile *Aquifex pyrophilus* at 1.9 Å resolution: structural basis for thermostability. J Mol Biol 1997;270:259–274.
4. Chang C, Park BC, Lee DS, Suh SW. Crystal structures of thermostable xylose isomerases from *Thermus caldophilus* and *Thermus thermophilus*: possible structural determinants of thermostability. J Mol Biol 1999;288:623–634.
5. Haney PJ, Badger JH, Buldak GL, Reich CI, Woese CR, Olsen GJ. Thermal adaptation analyzed by comparison of protein sequences from mesophilic and extremely thermophilic *Methanococcus* species. Proc Natl Acad Sci USA 1999;96:3578–3583.
6. Hasegawa J, Shimahara H, Mizutani M, Uchiyama S, Arai H, Ishii M, Kobayashi Y, Ferguson SJ, Sambongi Y, Igarashi Y. Stabilization of *Pseudomonas aeruginosa* cytochrome c(551) by systematic amino acid substitutions based on the structure of thermophilic *Hydrogenobacter thermophilus* cytochrome c(552). J Biol Chem 1999;274:37533–37537.
7. McDonald JH, Grasso AM, Rejto LK. Patterns of temperature adaptation in proteins from *Methanococcus* and *Bacillus*. Mol Biol Evol 1999;16:1785–1790.

8. Cambillau C, Claverie JM. Structural and genomic correlates of hyperthermostability. J Biol Chem 2000;275:32383–32386.
9. Das R, Gerstein M. The stability of thermophilic proteins: a study based on comprehensive genome comparison. Funct Integr Genomics 2000;1:76–88.
10. Declerck N, Machius M, Wiegand G, Huber R, Gaillardin C. Probing structural determinants specifying high thermostability in *Bacillus licheniformis* alpha-amylase. J Mol Biol 2000;301:1041–1057.
11. Kumar S, Tsai CJ, Nussinov R. Factors enhancing protein thermostability. Protein Eng 2000;13:179–191.
12. Perl D, Mueller U, Heinemann U, Schmid FX. Two exposed amino acid residues confer thermostability on a cold shock protein. Nat Struct Biol 2000;7:380–383.
13. Szilagyi A, Zavodszky P. Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits: results of a comprehensive survey. Struct Fold Des 2000;8:493–504.
14. Kreil DP, Ouzounis CA. Identification of thermophilic species by the amino acid compositions deduced from their genomes. Nucleic Acids Res 2001;29:1608–1615.
15. Kumar S, Sham YY, Tsai CJ, Nussinov R. Protein folding and function: the N-terminal fragment in adenylate kinase. Biophys J 2001;80:2439–2454.
16. Kumar S, Tsai CJ, Nussinov R. Thermodynamic differences among homologous thermophilic and mesophilic proteins. Biochemistry 2001;40:14152–14165.
17. Lehmann M, Wyss M. Engineering proteins for thermostability: the use of sequence alignments versus rational design and directed evolution. Curr Opin Biotechnol 2001;12:371–375.
18. Vieille C, Zeikus GJ. Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability. Microbiol Mol Biol Rev 2001;65:1–43.
19. Chakravarty S, Varadarajan R. Elucidation of factors responsible for enhanced thermal stability of proteins: a structural genomics based study. Biochemistry 2002;41:8152–8161.
20. Rosato V, Pucello N, Giuliano G. Evidence for cysteine clustering in thermophilic proteomes. Trends Genet 2002;18:278–281.
21. Criswell AR, Bae E, Stec B, Konisky J, Phillips GN Jr. Structures of thermophilic and mesophilic adenylate kinases from the genus *Methanococcus*. J Mol Biol 2003;330:1087–1099.
22. La D, Silver M, Edgar RC, Livesay DR. Using motif-based methods in multiple genome analyses: a case study comparing orthologous mesophilic and thermophilic proteins. Biochemistry 2003;42:8988–8998.
23. Machius M, Declerck N, Huber R, Wiegand G. Kinetic stabilization of *Bacillus licheniformis* alpha-amylase through introduction of hydrophobic residues at the surface. J Biol Chem 2003;278:11546–11553.
24. Chan CH, Liang HK, Hsiao NW, Ko MT, Lyu PC, Hwang JK. The relationship between local structural entropy and protein thermostability. Proteins 2004;57:684–691.
25. McDonald JH, Grasso AM, Rejto LK. Patterns of temperature adaptation in proteins from *Methanococcus* and *Bacillus*. Mol Biol Evol 1999;16:1785–1790.
26. Chakravarty S, Varadarajan R. Elucidation of determinants of protein stability through genome sequence analysis. FEBS Lett 2000;470:65–69.
27. La D, Silver M, Edgar RC, Livesay DR. Using motif-based methods in multiple genome analyses: a case study comparing orthologous mesophilic and thermophilic proteins. Biochemistry 2003;42:8988–8998.
28. Nakashima H, Fukuchi S, K. N. Compositional changes in RNA, DNA and proteins for bacterial adaptation to higher and lower temperatures. J Biochem 2003;133:507–513.
29. Neave HR, Worthington PL. Distribution-free tests. London; Unwin Hyman; 1988.
30. Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. Science 1997;278:631–637.