PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Using a strategy based on the concept of convergent evolution to identify residue substitutions responsible for thermal adaptation

Yeong-Shin Lin*

Institute of Bioinformatics, National Chiao Tung University, Hsinchu, Taiwan

ABSTRACT

Factors that are related to thermostability of proteins have been extensively studied in recent years, especially by comparing thermophiles and mesophiles. However, most of them are global characters. It is still not clear how to identify specific residues or fragments which may be more relevant to protein thermostability. Moreover, some of the differences among the thermophiles and mesophiles may be due to phylogenetic differences instead of thermal adaptation. To resolve these problems, I adopted a strategy to identify residue substitutions evolved convergently in thermophiles or mesophiles. These residues may therefore be responsible for thermal adaptation. Four classes of genomes were utilized in this study, including thermophilic archaea, mesophilic archaea, thermophilic bacteria, and mesophilic bacteria. For most clusters of orthologous groups (COGs) with sequences from all of these four classes of genomes, I can identify specific residues or fragments that may potentially be responsible for thermal adaptation. Functional or structural constraints (represented as sequence conservation) were suggested to have higher impact on thermal adaption than secondary structure or solvent accessibility does. I further compared thermophilic archaea and mesophilic bacteria, and found that the most diverged fragments may not necessarily correspond to the thermostability-determining ones. The usual approach to compare thermophiles and mesophiles without considering phylogenetic relationships may roughly identify sequence features contributing to thermostability; however, to specifically identify residue substitutions responsible for thermal adaptation, one should take sequence evolution into consideration.

Proteins 2008; 73:53–62.

Key words: thermostability; thermophile; mesophile; phylogeny; orthologue; protein structure; amino acid composition; function prediction.

INTRODUCTION

The nature of thermostable proteins has drawn much attention in the past few decades. Many protein sequence and structure features characteristic of thermophilic proteins were revealed by comparison to counterpart mesophilic homologues. One of the most noticeable features are differences in amino acid composition between the thermophilic and mesophilic proteins, especially charged versus polar (noncharged) residues.[1–9] This difference was further found to be preferentially located at the protein surface,[2,10–13] and the helices, especially the N-terminal residues.[11,14–16] Study showed that helices of thermophilic proteins appear to be more stable than those of the mesophilic homologues.[17] Other possible indicators of thermostability are asymmetrical substitution patterns for certain amino acid pairs[18]; amino acid coupling patterns[19]; local structural entropy[20]; tighter hydrophobic packing[6,21]; folds with high levels of contact trace[22,23]; a decrease in the entropy of unfolding[6]; fewer free cysteine amino acids except those involved in disulfide bridges and metal binding, or those inaccessible to the solvent[6,24,25]; and more side-chain-side-chain hydrogen bonds and ion pairs.[4,12,16,21,26]

However, despite the many factors described earlier that are related to thermostability of proteins, it is still not clear how to identify specific residues or fragments responsible for thermal adaptation, which may be useful for rational design[27] of thermal proteins. Many factors important for protein thermostability have varying and, sometimes, even opposing contributions to protein thermostability. For example, the position of amino acid changes in the structure may be even more important than a global increase in charged residues.[13,28] Moreover, each protein family seems to adopt its own strategy or mechanism to adapt to high or low temperatures.[4,6,29–31] Most importantly, not all differences among the thermophiles and mesophiles may be attributable to protein thermostability; some of them may be due to phylogenetic

Grant sponsor: National Science Council.
*Correspondence to: Yeong-Shin Lin, Institute of Bioinformatics, National Chiao Tung University, Hsinchu 30068, Taiwan. E-mail: yslin@faculty.nctu.edu.tw.
Received 8 September 2007; Revised 24 January 2008; Accepted 15 February 2008
Published online 2 April 2008 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/prot.22049

differences.[16,17,28,32,33] Most thermophiles are archaea, while most bacteria are mesophiles. The differentiation of amino acid compositions among phyla[34] suggests that it would be informative to account for phylogenetic relationships. In this study, I adopted a strategy to identify convergent evolution that allows us to recognize residue substitutions shared among thermophiles or mesophiles in a cluster of orthologous groups, which may be responsible for thermal adaptation.

MATERIAL AND METHODS

Dataset

Clusters of Orthologous Groups of proteins (COGs) were obtained from http://www.ncbi.nlm.nih.gov/COG/.[35] I selected 12 thermophilic archaea genomes (TA; *Archaeoglobus fulgidus*, *Methanococcus jannaschii*, *Sulfolobus solfataricus*, *Pyrococcus horikoshii*, *Pyrococcus abyssi*, *Methanopyrus kandleri AV19*, *Pyrobaculum aerophilum*, *Aeropyrum pernix*, *Methanothermobacter thermautotrophicus*, *Thermoplasma acidophilum*, *Thermoplasma volcanium*, *Halobacterium sp. NRC-1*); 1 mesophilic archaea genome (MA; *Methanosarcina acetivorans str.C2A*); 2 thermophilic bacteria genomes (TB; *Thermotoga maritima*, *Aquifex aeolicus*); and 7 mesophilic bacteria genomes (MB; *Nostoc sp. PCC 7120*, *Synechocystis sp.*, *Fusobacterium nucleatum*, *Treponema pallidum*, *Borrelia burgdorferi*, *Chlamydia trachomatis*, *Chlamydophila pneumoniae CWL029*). A COG may include more than one sequence (paralogue) from one genome. COGs with a PDB homologue and with at least one sequence from each of the above four classes were retained for further analyses.

Phylogenetic relationships

It is possible that distinct organisms may evolve similar traits independently as they both adapt to similar environments. This process is named as convergent evolution. In view of the idea that protein thermostability may evolve independently in archaea and bacteria, I used this concept to develop a strategy to identify residue substitutions responsible for thermal adaptation. The general phylogenetic relationships for archaea and bacteria are shown in Figure 1(A). For most homologous residues, species from the same lineage (archaea or bacteria) should share similar evolutionary history and should be clustered together [Fig. 1(B)]. However, for some residues that are responsible for thermal adaptation, that is, they are under strong selection, species living in similar environments (thermophiles or mesophiles) may be clustered together due to convergent evolution [Fig. 1(C)]. For this reason, I will define two types of tree topologies to describe the evolutionary relationships: PR tree [the species tree representing true phylogenetic relationships; Fig. 1(B)] and CE tree [a fictitious tree clustering species under convergent evolution together; Fig. 1(C)]. Although thermophilic or mesophilic origin of bacteria is still under debate,[36–41] PR tree and CE tree topologies hold for either case.

Amino acid composition as a character state

To identify residue substitutions responsible for thermal adaptation, I subdivided the protein sequences into residues or sliding windows and analyzed which tree topology (PR tree or CE tree) is preferentially supported by the substitutions occurred in the regions studied. Note that, thermal adaptation may not easily be achieved by single amino acid substitution at a specific site; in contrast, differences in amino acid compositions have been suggested to be the most noticeable characters differentiated between thermophiles and mesophiles.[1–9] I therefore examined the amino acid composition in a sliding window (the number of amino acids belonging to certain types, e.g., charged or polar) and coded this as the character state for each organism. Figure 1D shows the multiple sequences alignment for COG0020 as an example. The sliding window is on the original amino acid sequence; therefore gaps in the alignments have no influence and are not considered.

I first used, F_{cp}, the difference of the numbers of charged (Asp, Glu, Lys and Arg) and polar noncharged (Asn, Gln, Ser and Thr) amino acids[9] as a continuous character state for each organism to construct the parsimony tree (as described below). However, it should be noticed that Asp is not as preferred by thermophiles as the other three charged amino acids are and, in contrast, that Ala and Val are highly preferred by mesophiles and thermophiles, respectively.[9] I therefore defined another character state, F_{tm}, the number of Glu, Lys, Arg and Val versus the number of Asn, Gln, Ser, Thr, and Ala. To reveal whether these results could be obtained by chance, I permuted the amino acid compositions from the previous two character states as a new one, F_{permuted}, the number of Asp, Asn, Ser and Val versus the number of Glu, Gln, Thr, and Arg. I also used F_{aromatic}, the number of aromatic residues (Phe, Trp and Tyr), and subdivided F_{tm} into $F_{\mathrm{tm\text{-}c}}$ (charged: Glu, Lys, and Arg), $F_{\mathrm{tm\text{-}p}}$ (polar: Asn, Gln, Ser, and Thr) and $F_{\mathrm{tm\text{-}np}}$ (nonpolar: Val vs. Ala) for extensive analyses.

To obtain an accurate measure of amino acid composition, a sufficient window size is required; on the other hand, when the window size is large, the selection pressure on the fragment may be diluted, and the difference of selection pressures between fragments may therefore be blurred. These two issues should be balanced. I used 15 residues as the windows size in this study after preliminary examinations. Slightly increasing or decreasing the window size does not change the results.

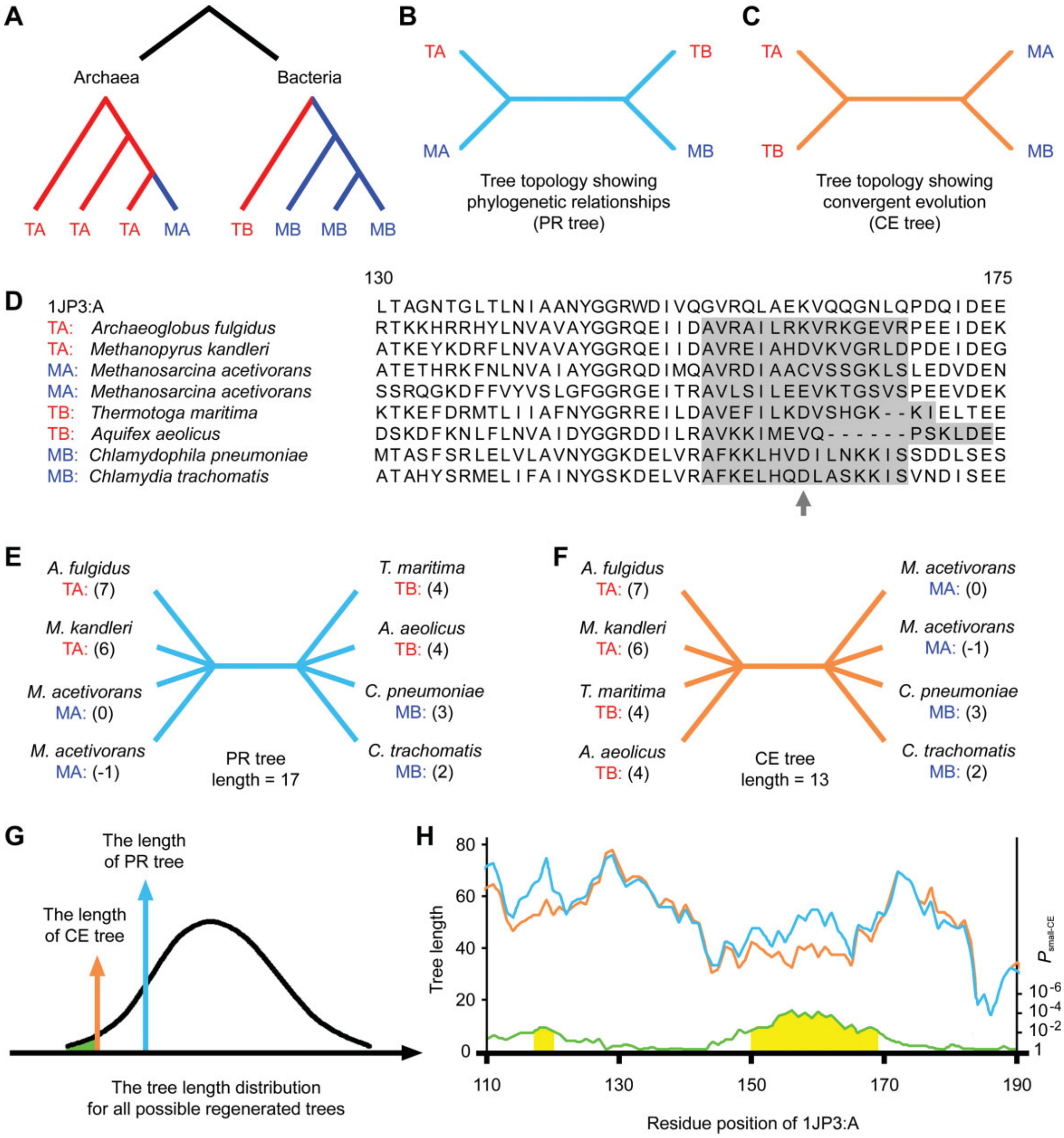
A
Archaea
Bacteria
TA TA TA MA TB MB MB MB
B
TA
TB
MA
MB
Tree topology showing phylogenetic relationships (PR tree)
C
TA
MA
TB
MB
Tree topology showing convergent evolution (CE tree)
D
130
175
1JP3:A LTAGNTGLTLNIAANYGGRWDIVQGVRQLAEKVQQGNLQPDQIDEE
TA: Archaeoglobus fulgidus RTKKHRRHYLNVAVAYGGRQEIIDAVRAILRKVRKGEVRPEEIDEK
TA: Methanopyrus kandleri ATKEYKDRFLNVAVAYGGRQEIIDAVREIAHDVKVGRLDPDEIDEG
MA: Methanosarcina acetivorans ATETHRKFNLNVAIAYGGRQDIMQAVRDIAACVSSGKLSLEDVDEN
MA: Methanosarcina acetivorans SSRQGKDFFVYVSLGFGGRGEITRAVLSILEEVKTGSVSPEEVDEK
TB: Thermotoga maritima KTKEFDRMTLIIAFNYGGRREILDAVEFILKDVSHGK--KIELTEE
TB: Aquifex aeolicus DSKDFKNLFLNVAIDYGGRDDILRAVKKIMEVQ------PSKLDEE
MB: Chlamydophila pneumoniae MTASFSRLELVLAVNYGGKDELVRAFKKLHVDILNKKISSDDLSES
MB: Chlamydia trachomatis ATAHYSRMELIFAINYGSKDELVRAFKELHQDLASKKISVNDISEE
E
A. fulgidus TA: (7)
M. kandleri TA: (6)
M. acetivorans MA: (0)
M. acetivorans MA: (-1)
T. maritima TB: (4)
A. aeolicus TB: (4)
C. pneumoniae MB: (3)
C. trachomatis MB: (2)
PR tree length = 17
F
A. fulgidus TA: (7)
M. kandleri TA: (6)
T. maritima TB: (4)
A. aeolicus TB: (4)
M. acetivorans MA: (0)
M. acetivorans MA: (-1)
C. pneumoniae MB: (3)
C. trachomatis MB: (2)
CE tree length = 13
G
The length of PR tree
The length of CE tree
The tree length distribution for all possible regenerated trees
H
Tree length
80
60
40
20
0
110
130
150
170
190
Residue position of 1JP3:A
$P_{small\text{-}CE}$
10^{-6}
10^{-4}
10^{-2}
1

Figure 1

*The strategy based on the concept of convergent evolution developed in this study. (**A**) The general phylogenetic tree. Red and blue indicate thermophiles and mesophiles, respectively. TA, thermophilic archaea; MA, mesophilic archaea; TB, thermophilic bacteria; MB, mesophilic bacteria. (**B**) In the tree topology showing true phylogenetic relationships (PR tree, represented as cyan), archaea and bacteria are separately clustered. (**C**) In the fictitious tree topology clustering species under convergent evolution together (CE tree, represented as orange), thermophilics and mesophilics are separately clustered. (**D**) A partial alignment of amino acid sequences collected in COG0020, undecaprenyl pyrophosphate synthase. Only 9 out of 28 sequences are displayed in this figure, including one PDB structure (1JP3:A). The numbers represent the corresponding residue positions of 1JP3:A. The arrow indicates the aligned column position corresponding to the 161st residue of 1JP3:A. Amino acids located in a 15 aa-sliding window centered on the column residues are displayed with the gray shadow. (**E**) Using the difference of the numbers of charged and polar residues (F_{cp}) in the sliding window showing in Figure 1(D) as the character state (the numbers shown in the parentheses), the length of PR tree (the summation of the changes between the numbers) is calculated as 17 based on the parsimony principle, that is assigning 3 and 3, or 4 and 4 for the two internal nodes. This is the minimized tree length. (**F**) Similarly, the length of CE tree is calculated as 14 (with 4 and 2 for the two internal nodes). (**G**) A hypothetical distribution of tree lengths for trees randomly regenerated from the PR tree. The green area indicates the probability that a randomly regenerated tree has a tree length equal to or shorter than the length of CE tree ($P_{small\text{-}CE}$). (**H**) For each aligned column of COG0020, the tree lengths for PR tree and CE tree are calculated using all the 27 sequences in COG0020 (excluding 1JP3:A), and represented as cyan and orange lines, respectively. The $P_{small\text{-}CE}$ value is represented as the green line. The column positions with $P_{small\text{-}CE} < 10^{-2}$ (thermal adaptation sites) are represented as yellow.*

Calculation of the tree length

For each aligned column, the character state (of the sliding window centered on the column residue) for each protein sequence was assigned to each branch in the two models [PR tree and CE tree; Fig. 1(E,F)]. The tree length for both models for each aligned column can therefore be calculated by summing all character state changes between nodes on a tree. To simplify the calculation, I assumed a topology where all species radiate (i.e., form a polytomy) from one of the two internal nodes. Parsimony was used to calculate the tree length, that is, the character states of the two internal nodes were reconstructed to minimize the tree length [e.g., Fig. 1(E,F)].

Assessing statistical significance

For an aligned column where CE tree has a smaller tree length than PR tree, residues surrounding the column (in the sliding window) are probably under convergent evolution, which means they might be responsible for thermal adaptation. To clarify whether the small length of CE tree can be derived by chance, I assumed PR tree is the real tree topology and regenerated various trees from PR tree to construct a distribution of tree length for all possible trees. It should be noticed that, the difference between PR tree and CE tree is the switching between MA and TB sequences [Fig. 1(E,F)]. Assuming there are n_1 MA sequences out of m_1 archaea sequences, and n_2 TB sequences out of m_2 bacteria sequences, a regenerated tree is performed by switching n_1 archaea sequences and n_2 bacteria sequences from PR tree. The total number of possible regenerated trees is $C_{n_1}^{m_1} \times C_{n_2}^{m_2}$. The distribution of tree length for all possible regenerated trees was thus obtained [Fig. 1(G)], unless the number of possible regenerated trees is larger than 10^5. In that case, to save the computational time, a switching between n_1 randomly selected archaea sequences and n_2 randomly selected bacteria sequences from PR tree was repeated 10^5 times. The probability ($P_{\text{small-CE}}$) that a randomly regenerated tree has a tree length equal to or smaller than the length of CE tree can thus be calculated [Fig. 1(G)]. I therefore defined the aligned column with its CE tree length significantly small ($P_{\text{small-CE}} < 10^{-2}$) as the "thermal adaptation site." It should be noticed that not all substitutions causing thermal adaptation could be detected using this method, especially when the number of thermal adaptation substitutions is much smaller than that of the background mutations ($P_{\text{small-CE}}$ may not be significantly small). COGs with their total number of possible regenerated trees less than 100 were discarded because their sample sizes were too small to have the power to identify thermal adaptation sites. Figure 1(H) is an example showing $P_{\text{small-CE}}$ and the lengths of PR tree and CE tree using F_{cp} as the character state for each aligned column of COG0020.

Protein structures

The secondary structure assignment and the solvent accessible surface areas (ACC) for each residue of the PDB homologues were obtained from DSSP (database of secondary structure assignments for all protein entries in the PDB; http://swift.cmbi.ru.nl/gv/dssp/).[42] The DSSP method defines eight secondary structures according to their hydrogen bonding patterns: α-helix (H), 3_{10}-helix (G), π-helix (I), extended β-strand (E), isolated β-strand (B), turn (T), bend (S), and coil (U). Relative solvent accessibility (RelACC) was the ACC for each residue divided by the maximum value of ACC for the amino acid (represented in percentage), which is estimated from a Gly-X-Gly extended tripeptide conformation.

The secondary structure and RelACC value for each residue of the PDB homologue for each COG were compiled. Residues in a sliding window centered on a thermal adaptation site were annotated. For each category subdivided based on the secondary structures and RelACC values (secondary structures G, I, and B were ignored due to their extremely small sample sizes), the probability that a residue locates in a sliding window centered on a thermal adaptation site (P_{thermal}) can thus be estimated. Residues in a category with a high P_{thermal} are suggested to have higher chance to be potentially responsible, or at least to have higher chance to be detected as being responsible for thermal adaptation using the proposed method.

Sequence conservation

Considering the sample size, conservations in a COG were only estimated using TA and MB sequences. The conservation at one aligned column was estimated using sequence entropy, $S = -\sum_{j=1}^{20} \pi_j \ln \pi_j$, where π_j is the frequency of each amino acid in this column. The two conservation values of thermophilic archaea and mesophilic bacteria, S_{TA} and S_{MB}, were calculated separately and then averaged as S for each aligned column in a COG. Columns with $S < 1$ were defined as conserved sites. The proportion of conserved sites (f) in a sliding window was thus calculated.

RESULTS AND DISCUSSION

Protein structure and thermal adaptation

I subdivided residues in all PDB homologues based on their secondary structure assignments and solvent accessibility to investigate the tendency of these residues potentially being responsible for thermal adaptation (see Fig. 2). Consistent with previous studies,[2,10–16] the result shows that helix residues and exposed resides are important for thermal adaptation using F_{cp} as the character state to perform the analysis. The P_{thermal} value for α-helix is significantly higher than each of the other four

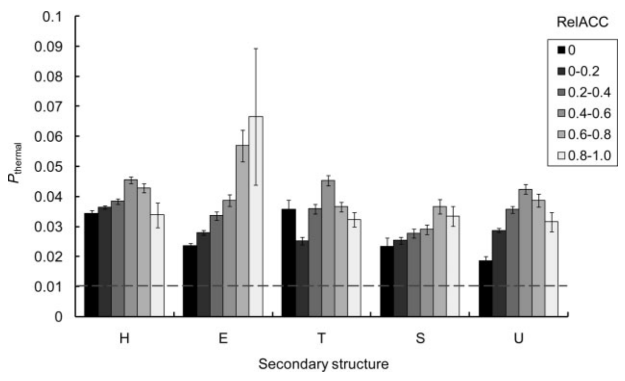

Figure 2

The relationships between secondary structure, solvent accessibility, and $P_{thermal}$ revealed using F_{cp} as the character state. The error bars represent standard error. The gray dotted line indicates background $P_{thermal}$ = 0.01.

secondary structures (χ^2 test, $P < 10^{-4}$). Residues with RelACC 0.4~0.6 also have their P_{thermal} value significantly higher than each category with RelACC less than 0.4 (χ^2 test, $P < 10^{-16}$). However, the extremely exposed residues are not the most preferred ones as expected. When RelACC is increased from 0.4~0.6 to 0.6~0.8, P_{thermal} is slightly decreased; when RelACC is further increased to 0.8~1, P_{thermal} is significantly reduced (χ^2 test, $P < 10^{-6}$).

It was suggested that introducing stabilization mutations on protein surface may stabilize the reversibly unfolded state without creating volume interferences in the core.[6] This study further indicates that mutations at the extremely exposed regions may only provide limited contributions to protein thermostability.

Using different amino acid compositions as the character states

Although the difference of the numbers of charged and polar amino acids (F_{cp}) has long been recognized as an important difference between thermophiles and mesophiles,[1–9] Figure 3(A) shows that more thermal adaptation sites were recognized using F_{tm} as the character state instead. The relationships between protein structure and thermal adaptation are generally consistent with that obtained using F_{cp} (see Fig. 2). To illustrate whether these results could be derived by chance, F_{permuted} was used as a control character state. Each secondary structure and RelACC category shows background P_{thermal} consistently, that is, 0.01 [Fig. 3(B)], because $P_{\text{small-CE}} < 10^{-2}$ was used as the criteria to define thermal adaptation site. Although aromatic clusters on protein surface was suggested to correlate with thermostability,[43] using F_{aromatic} as the character state also only obtains background P_{thermal} [Fig. 3(C)]. This result implies that either the aromatic clustering is not a general feature for thermostability, or the clustering is not derived by increasing the number of aromatic amino acids but just special rearrangements.

I further subdivided F_{tm} into charged amino acids ($F_{\text{tm-c}}$), polar amino acids ($F_{\text{tm-p}}$), and nonpolar amino acids ($F_{\text{tm-np}}$) to compare their contributions to thermal adaptation. The obtained results indicate that the increasing of charged residues [Fig. 3(D)] has much higher impact than the decreasing of polar residues [Fig. 3(E)] for thermostability. The charged residue substitutions dominate at the helices (χ^2 test, $P < 10^{-8}$). It is likely that the increased charged residues usually locate around the thermal adaptation sites; in contrast, the replacements of polar residues may distribute along the whole protein more dispersedly. Except for the exposed helices, P_{thermal} values estimated using $F_{\text{tm-np}}$ as the character state are only slightly higher than the background value [Fig. 3(F)], which suggests that Ala and Val substitutions may also tend to occur globally.

Sequence conservation and thermal adaptation

It has been proposed that stabilizing interactions in thermophilic proteins tend to locate in the less conserved areas of the protein.[6] I subdivided aligned columns of each COG based on their sequence conservation, that is, the proportion of conserved sites (f) in the sliding window centered on the column. It was found that when most residues in the local environment are conserved, their P_{thermal} values are consistently low regard less their secondary structures and solvent accessibility; however, they are still higher than the background value, 0.01 [Fig. 4(A)]. Because exposed residues are usually less conserved,[44] residues with high RelACC have comparatively high variance due to their small sample size. Note that I used TA sequences and MB sequences independently to estimate the conservation. Although conserved mutations between thermophilic and mesophilic subfamilies (conserved intra subfamily but diverged between subfamilies) were suggested to provide key residue differences that are potentially related to increased thermostability,[7] these regions should have high f values, and may therefore not necessarily relate to thermal adaptation based on results in this study.

When the number of nonconserved sites in the window increases, P_{thermal} increases, especially for helix or moderately exposed residues [Fig. 4(B,C)]. When most sites in the window are nonconserved, each category shows high P_{thermal} [Fig. 4(D)].

Most residues in a protein are under constraints to maintain protein function or structural stability. Mutations of residues under weaker constraints may therefore have higher chance to obtain extra thermostability, that is, they are more evolvable. The differentiation of P_{thermal}

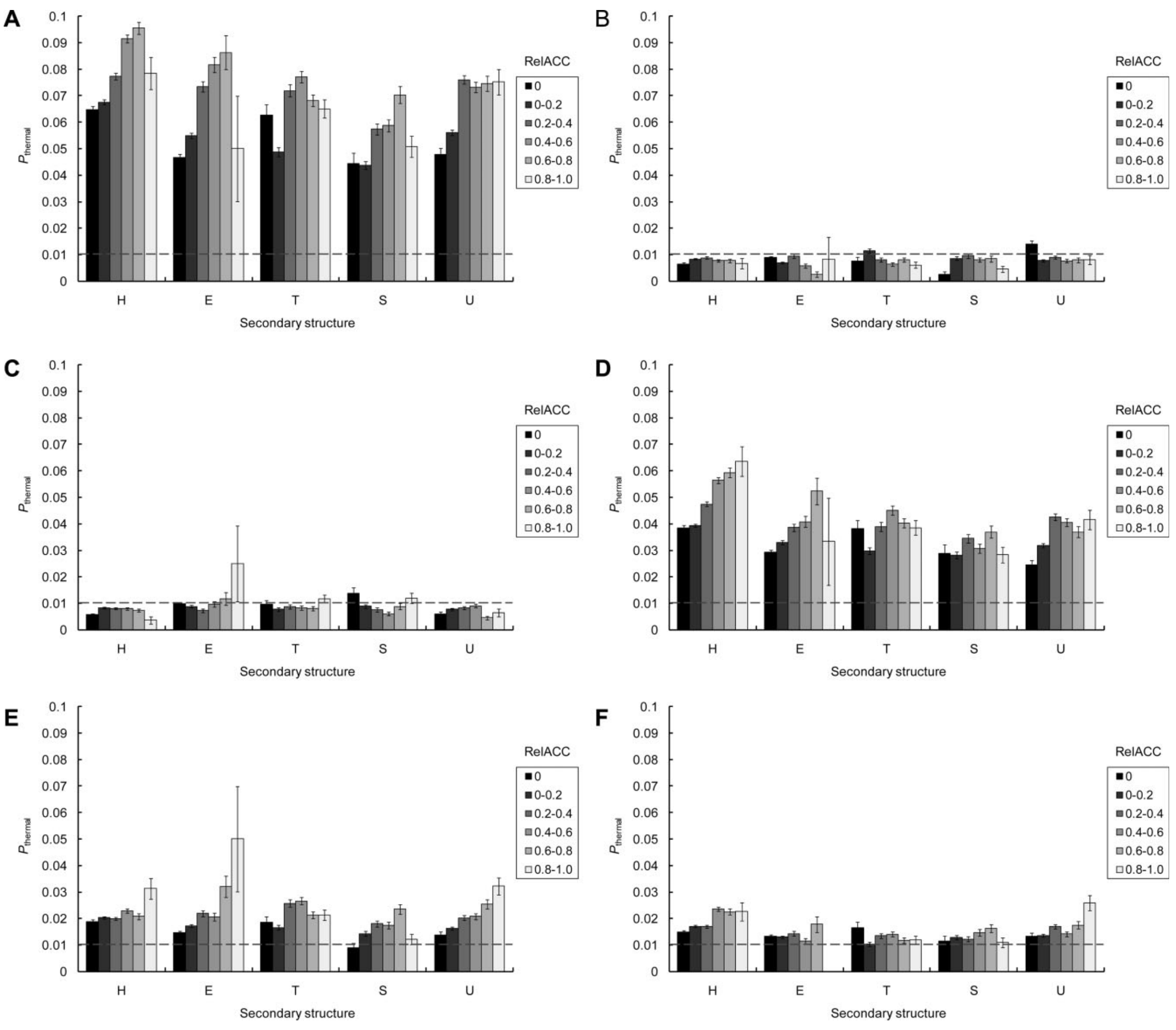

Figure 3

The relationships between secondary structure, solvent accessibility, and $P_{thermal}$ revealed using (**A**) F_{tm}; (**B**) $F_{permuted}$; (**C**) $F_{aromatic}$; (**D**) $F_{tm\text{-}c}$; (**E**) $F_{tm\text{-}p}$; (**F**) $F_{tm\text{-}np}$ *as the character states. The error bars represent standard error. The gray dotted line indicates background* $P_{thermal} = 0.01$.

among different secondary structures or regions with different solvent accessibility is only effective under certain selection pressures. Note that although exposed residues tend to be nonconserved and have high P_{thermal}, nonconserved buried residues have significantly higher P_{thermal} than conserved exposed residues ($f \leq 0.25$ with RelACC $= 0$ vs. $f > 0.75$ with RelACC >0.4, χ^2 test, $P < 10^{-8}$). Based on these results, I suggest that the chance that a residue is important for thermal adaptation correlates with protein secondary structure, solvent accessibility, and functional or structural constraints (represented as sequence conservation) independently, where functional or structural constraints may have the most significant impact.

Thermal adaptation versus phylogenetic differences

To investigate whether my strategy can distinguish specific substitutions responsible for thermal adaptation from the phylogenetic differences, I used TA and MB sequences to calculate the differentiation between thermophiles and mesophiles. These two groups of species compose the majority of thermophiles and mesophiles, respectively. For each aligned column, the averaged character state values (e.g., F_{tm}) for TA sequences and MB sequences were calculated separately. The absolute value of their difference, d, was thus used to represent the differentiation, which may therefore include the differences attributed to both thermostability and phylogeny.

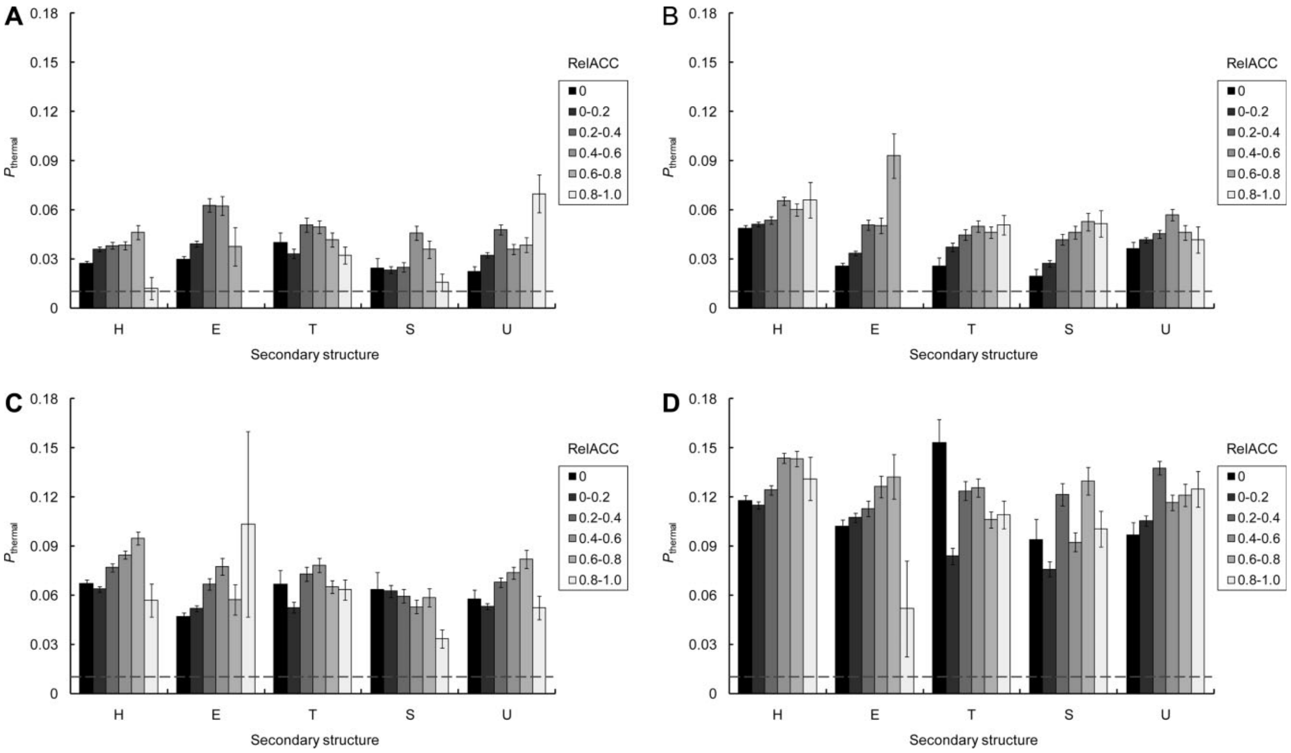

Figure 4

*The relationships between secondary structure, solvent accessibility, and $P_{thermal}$ revealed using F_{tm} as the character state with (**A**) $f > 0.75$; (**B**) $0.75 \geq f > 0.5$; (**C**) $0.5 \geq f > 0.25$ (**D**) $f \leq 0.25$; where f is the proportion of conserved sites in a sliding window. The error bars represent standard error. The gray dotted line indicates background $P_{thermal} = 0.01$.*

Residues with low $P_{\text{small-CE}}$ values (thermal adaptation sites) usually have high d values (they may be responsible for thermal adaptation); in contrast, residues with high d values may not necessarily have their $P_{\text{small-CE}}$ values significantly low (some of them may be attributed to phylogenetic differences). The result shows that residues with the most significant $P_{\text{small-CE}}$ values and residues with the highest d values can be nonoverlapped (e.g., COG0057 in Fig. 5). COG0057 is a protein family of glyceraldehyde-3-phosphate dehydrogenase (GAPDH). Five fragments in the 3D structure differ significantly between archaea and bacteria.[45] Fragment I (residues 39-47 on 1B7G:O) is an α-helix corresponding to the pick with very low $P_{\text{small-CE}}$ values (residues 41-49). Fragment II (56–78) includes two α-helices existing in archaea but not in bacteria; where parts of this region are also recognized as thermal adaptation sites (52–54, 62, and 65). These two regions may correlate with thermal adaptation based on results in this study. Fragment III (169–178) is the so-called S-loop. A part of the S-loop and its upstream β-sheet have very high d values, but with insignificant $P_{\text{small-CE}}$ values. (160–173) Similar situation is applied for the downstream region (273–286) of fragment IV (257–268) which includes another extra α-helix in archaea structure. The differentiation of these two regions may therefore be due to phylogeny. Fragment V (320–332) is the C terminal α-helix only existing in archaea, and is not analyzed here.

It is likely that those phylogenetic differences in GAPDH are mainly due to a relocation of the active-site residues within the enzyme's catalytic domain between archaea and bacteria.[45] Most regions neighbor to the active-site residues (Ser138, Cys139, Asn140, Arg166, Arg167, His192, His193 and His219) show such phylogenetic differences (136–142, 160–173, corresponding to the upstream of fragment III, and 188–191). The binding-site residues of the $NADP^+$ adenosine 2′ phosphate (Lys33, Thr34 and Ser35) also have high d values and insignificant $P_{\text{small-CE}}$ values.

Using the whole dataset to reveal the relationship between the significance of $P_{\text{small-CE}}$ and the difference, d, we can find that many aligned columns show either high significance of $P_{\text{small-CE}}$ or high difference, d (see Fig. 6), although a positive correlation between them still can be recognized ($R = 0.503$). This result suggests that comparing thermophiles and mesophiles without considering phylogenetic relationships can roughly identify the characters contributing to thermostability. However, to specifically identify residue substitutions responsible for ther-

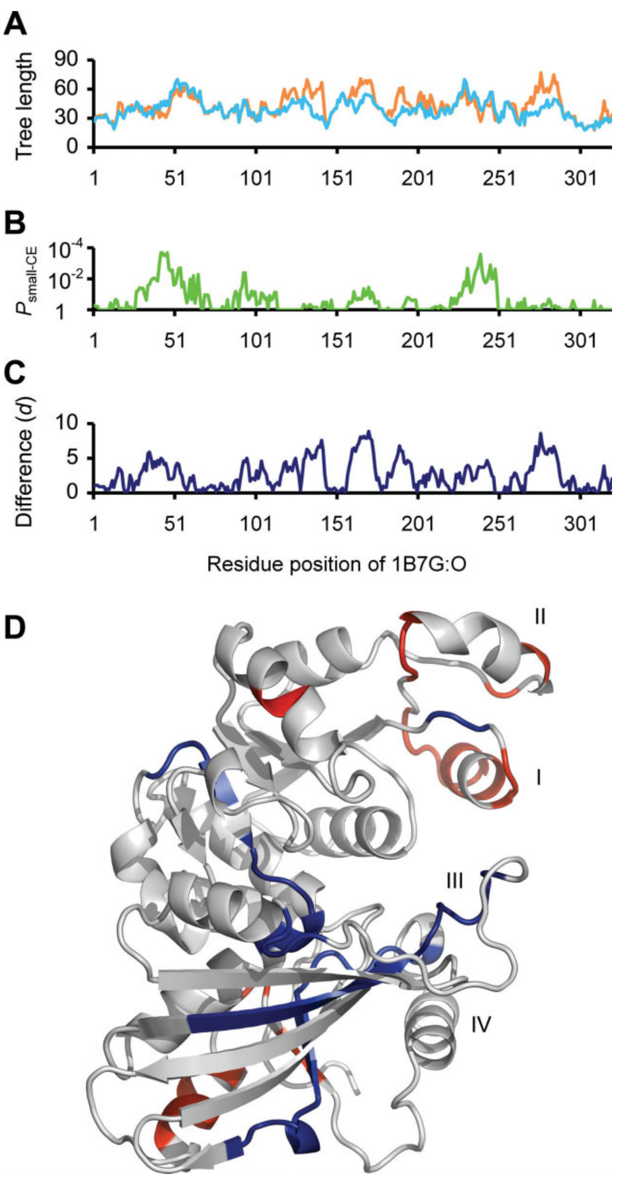

Figure 5

*Using F_{tm} as the character state to analyze COG0057 (similar results were obtained using F_{cp} as the character state, data not shown). (**A**) The tree lengths for PR tree and CE tree represented as cyan and orange lines, respectively. (**B**) The $P_{small\text{-}CE}$ value. (**C**) The absolute value of the difference (d) between the mean of TA sequences and the mean of MB sequences. (**D**) The 3D structure of 1B7G:O. Thermal adaptation sites ($P_{small\text{-}CE} < 10^{-2}$) are denoted as red, while residues with d > 5 are denoted as blue. Fragments I–IV are also indicated.*

mal adaptation, using the strategy proposed here may be more convincing.

SUMMARY

The strategy based on the concept of convergent evolution developed in this study successfully identified particular fragments potentially being responsible for thermal adaptation for most COGs (with at least couple sequences from each of the four classes) despite that the identification power of this strategy may be limited when the number of sequences in a COG is small. Although different organisms may potentially utilize different mechanisms to adapt thermal environments,[23,28,31] the mutual features shared among thermophiles or mesophiles still can be identified through this strategy. Unexpected gene exchange between members of archaea and bacteria were reported.[46–49] However, that happens infrequently, and most proteome should be developed to adapt the environment by the organisms themselves.[23] Nevertheless, such gene exchange may misinterpret most regions in a protein as thermal adaptation sites, which was not seen in the analyzed dataset. The thermal adaptation sites identified here should therefore be ideal candidates for rational design of thermal proteins.

For a given protein without orthologous information, the general trends summarized from this study can provide some suggestions to improve thermostability, for example the addition of charged residues at the nonconserved regions, or at the weakly conserved and moderately exposed helices. However, even for these regions, the probability of a residue to locate surrounding thermal adaptation sites (being responsible for thermal adaptation) is still hardly more than 10%. Before a more general feature is discovered, using orthologous sequences and the strategy proposed in this study to identify thermal adaptation sites is still the most reliable way.

It is worth to notice that this strategy is not restricted to identify thermal adaptation. Residues responsible for protein functions or features evolved independently multiple times in different organisms could also be identified

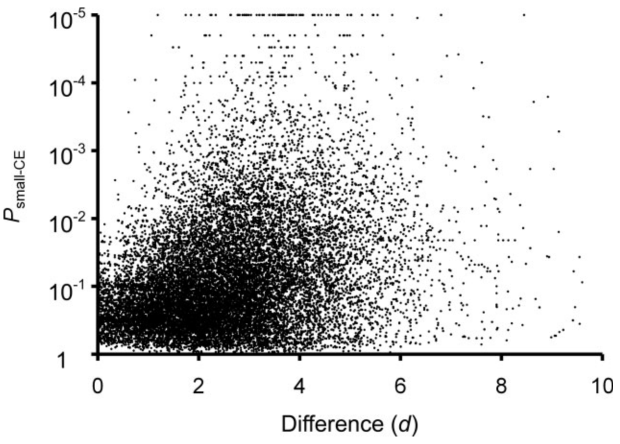

Figure 6

The relationships between $P_{small\text{-}CE}$ (the probability that a randomly regenerated tree has a tree length equal to or smaller than the length of CE tree) and the absolute difference value (d) between the mean of TA sequences and the mean of MB sequences using F_{tm} as the character state.

using it. With more and more genomes been resolved, this convergent evolution-aided strategy should become a useful tool for proteomic analyses.

ACKNOWLEDGMENTS

The author thank Joshua Rest and the anonymous reviewer for the valuable comments, and the Structural Bioinformatics Core Facility at National Chiao Tung University for the hardware and software support.

REFERENCES

1. 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.
2. Cambillau C, Claverie J-M. Structural and genomic correlates of hyperthermostability. J Biol Chem 2000;275:32383–32386.
3. Chakravarty S, Varadarajan R. Elucidation of determinants of protein stability through genome sequence analysis. FEBS Lett 2000; 470:65–69.
4. Szilagyi A, Zavodszky P. Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits: results of a comprehensive survey. Structure 2000;8:493–504.
5. Kreil DP, Ouzounis CA. Identification of thermophilic species by the amino acid compositions deduced from their genomes. Nucleic Acids Res 2001;29:1608–1615.
6. Vieille C, Zeikus G. Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability. Microbiol Mol Biol Rev 2001;65:1–43.
7. 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.
8. Nakashima H, Fukuchi S, Nishikawa K. Compositional changes in RNA, DNA and proteins for bacterial adaptation to higher and lower temperatures J Biochem 2003;133:507–513.
9. Suhre K, Claverie J-M. Genomic correlates of hyperthermostability, an update. J Biol Chem 2003;278:17198–17202.
10. Fukuchi S, Nishikawa K. Protein surface amino acid compositions distinctively differ between thermophilic and mesophilic bacteria. J Mol Biol 2001;309:835–843.
11. Chakravarty S, Varadarajan R. Elucidation of factors responsible for enhanced thermal stability of proteins: a structural genomics based study. Biochemistry 2002;41:8152–8161.
12. Alsop E, Silver M, Livesay DR. Optimized electrostatic surfaces parallel increased thermostability: a structural bioinformatic analysis. Protein Eng 2003;16:871–874.
13. Thorvaldsen S, Hjerde E, Fenton C, Willassen NP. Molecular characterization of cold adaptation based on ortholog protein sequences from *Vibrionaceae species.* Extremophiles 2007;11:719–732.
14. Querol E, Perez-Pons JA, Mozo-Villarias A. Analysis of protein conformational characteristics related to thermostability. Protein Eng 1996;9:265–271.
15. Das R, Gerstein M. The stability of thermophilic proteins: a study based on comprehensive genome comparison. Funct Integr Genom 2000;1:76–88.
16. Kumar S, Tsai C-J, Nussinov R. Factors enhancing protein thermostability. Protein Eng 2000;13:179–191.
17. Facchiano AM, Colonna G, Ragone R. Helix stabilizing factors and stabilization of thermophilic proteins: an X-ray based study. Protein Eng 1998;11:753–760.
18. McDonald JH, Grasso AM, Rejto LK. Patterns of temperature adaptation in proteins from *Methanococcus* and *Bacillus.* Mol Biol Evol 1999;16:1785–1790.
19. Liang H-K, Huang C-M, Ko M-T, Hwang J-K. Amino acid coupling patterns in thermophilic proteins. Proteins 2005;59:58–63.
20. Chan C-H, Liang H-K, Hsiao N-W, Ko M-T, Lyu P-C, Hwang J-K. Relationship between local structural entropy and protein thermostability. Proteins 2004;57:684–691.
21. 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.
22. England JL, Shakhnovich BE, Shakhnovich EI. Natural selection of more designable folds: a mechanism for thermophilic adaptation. Proc Natl Acad Sci USA 2003;100:8727–8731.
23. Berezovsky IN, Shakhnovich EI. Physics and evolution of thermophilic adaptation. Proc Natl Acad Sci USA 2005;102:12742–12747.
24. Rosato V, Pucello N, Giuliano G. Evidence for cysteine clustering in thermophilic proteomes. Trends Genet 2002;18:278–281.
25. Schneider D, Liu Y, Gerstein M, Engelman DM. Thermostability of membrane protein helix-helix interaction elucidated by statistical analysis. FEBS Lett 2002;532:231–236.
26. Vogt G, Woell S, Argos P. Protein thermal stability, hydrogen bonds, and ion pairs. J Mol Biol 1997;269:631–643.
27. Hellinga HW. Rational protein design: combining theory and experiment. Proc Natl Acad Sci USA 1997;94:10015–10017.
28. Robinson-Rechavi M, Alibes A, Godzik A. Contribution of electrostatic interactions, compactness and quaternary structure to protein thermostability: lessons from structural genomics of *Thermotoga maritima.*. J Mol Biol 2006;356:547–557.
29. Jaenicke R. Stability and stabilization of globular proteins in solution. J Biotechnol 2000;79:193–203.
30. Gianese G, Bossa F, Pascarella S. Comparative structural analysis of psychrophilic and meso- and thermophilic enzymes. Proteins 2002; 47:236–249.
31. Beeby M, O'Connor BD, Ryttersgaard CB, Daniel R, Perry LJ, Yeates TO. The genomics of disulfide bonding and protein stabilization in thermophiles PLoS Biol 2005;3:e309.
32. Arnold FH, Wintrode PL, Miyazaki K, Gershenson A. How enzymes adapt: lessons from directed evolution. Trends Biochem Sci 2001;26: 100–106.
33. McDonald JH. Patterns of temperature adaptation in proteins from the bacteria *Deinococcus radiodurans* and *Thermus thermophilus.* Mol Biol Evol 2001;18:741–749.
34. Pe'er I, Felder CE, Man O, Silman I, Sussman JL, Beckmann JS. Proteomic signatures: amino acid and oligopeptide compositions differentiate among phyla. Proteins 2004;54:20–40.
35. Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. Science 1997;278:631–637.
36. Barns SM, Delwiche CF, Palmer JD, Pace NR. Perspectives on archaeal diversity, thermophily and monophyly from environmental rRNA sequences. Proc Natl Acad Sci USA 1996;93:9188–9193.
37. Galtier N, Tourasse N, Gouy M. A nonhyperthermophilic common ancestor to extant life forms. Science 1999;283:220–221.
38. Bocchetta M, Gribaldo S, Sanangelantoni A, Cammarano P. Phylogenetic depth of the bacterial genera *Aquifex* and *Thermotoga* inferred from analysis of ribosomal protein, elongation factor, and RNA polymerase subunit sequences. J Mol Evol 2000;50:366–380.
39. Daubin V, Gouy M, Perriere G. A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. Genome Res 2002;12:1080–1090.
40. Gaucher EA, Thomson JM, Burgan MF, Benner SA. Inferring the palaeoenvironment of ancient bacteria on the basis of resurrected proteins. Nature 2003;425:285–288.
41. Iwabata H, Watanabe K, Ohkuri T, Yokobori S-I, Yamagishi A. Thermostability of ancestral mutants of *Caldococcus noboribetus* isocitrate dehydrogenase. FEMS Microbiol Lett 2005;243:393–398.

42. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
43. Kannan N, Vishveshwara S. Aromatic clusters: a determinant of thermal stability of thermophilic proteins. Protein Eng 2000;13:753–761.
44. Bustamante CD, Townsend JP, Hartl DL. Solvent accessibility and purifying selection within proteins of *Escherichia coli* and *Salmonella enterica*. Mol Biol Evol 2000;17:301–308.
45. Isupov MN, Fleming TM, Dalby AR, Crowhurst GS, Bourne PC, Littlechild JA. Crystal structure of the glyceraldehyde-3-phosphate dehydrogenase from the hyperthermophilic archaeon *Sulfolobus solfataricus*. J Mol Biol 1999;291:651–660.
46. Ibba M, Bono JL, Rosa PA, Soll D. Archaeal-type lysyl-tRNA synthetase in the Lyme disease spirochete *Borrelia burgdorferi*. Proc Natl Acad Sci USA 1997;94:14383–14388.
47. Klenk H-P, Clayton RA, Tomb J-F, White O, Nelson KE, Ketchum KA, Dodson RJ, Gwinn M, Hickey EK, Peterson JD, Richardson DL, Kerlavage AR, Graham DE, Kyrpides NC, Venter JC. The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. Nature 1997;390:364–370.
48. Deckert G, Warren PV, Gaasterland T, Young WG, Lenox AL, Graham DE, Overbeek R, Snead MA, Keller M, Aujay M, Huber R, Feldman RA, Short JM, Olsen GJ, Swanson RV. The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. Nature 1998;392:353–358.
49. Pennisi E. Genome data shake tree of life. Science 1998;280:672–674.