Protein Thermostabilizing Factors: High Relative Occurrence of Amino Acids, Residual Properties, and Secondary Structure Type in Different Residual State

Seung Pil Pack · Taek Jin Kang · Young Je Yoo

Abstract The relative occurrences of amino acids, residual properties, and secondary structure type found in the residual structure states were compared between thermophilic and mesophilic proteins to find out the protein-thermostabilizing factors. The thermostabilizing patterns in each residual structure state are as follows: (1) in fully exposed state, higher relative occurrences of GLN, ILE, and PHE; (2) in exposed state, higher relative occurrences of ARG, GLU, salt bridges, the residue with low solvation energy, and the residues in 3/10 helix, and lower relative occurrences of ALA, SER, and VAL; (3) in partially exposed state, higher relative occurrence of flexible residue and lower relative occurrence of SER; (4) in buried state, higher relative occurrences of ARG and GLU, and lower relative occurrence of MET; and (5) in well-buried state, higher relative occurrences of ALA, cation–pi interaction, the residues in 3/10 helix, and lower relative occurrences of ASP, GLY, and the residues in the extended beta strand. These findings could be useful for developing protein thermostabilization strategies according to each residual structure state.

Keywords Thermostabilizing factors · Residual structure state · Systematical analysis · Thermophilic protein

---

Seung Pil Pack and Taek Jin Kang equally contributed to this paper.

S. P. Pack (✉)
Department of Biotechnology and Bioinformatics, Korea University,
Sejong 339-700, South Korea
e-mail: spack@korea.ac.kr

T. J. Kang
Department of Chemical and Biochemical Engineering, Dongguk University,
Seoul 100-715, South Korea

Y. J. Yoo (✉)
School of Chemical and Biological Engineering, Seoul National University, Seoul 151-742, South Korea
e-mail: yjyoo@snu.ac.kr

Introduction

The major goal of protein engineering is to design novel proteins and to develop protein-based bioprocess required for production of high-value biomolecules in various areas of pharmaceutical, agricultural, industrial, and environmental fields [1]. For industrial application, the engineered protein or enzyme should have highly stable reactivity even under harsh environments such as high temperature, salinity, and acidity [2]. To develop novel proteins for its industrial application via a rational design strategy, the molecule-level mechanisms of protein thermostability should be revealed and understood [3, 4]. As one research approach, analysis of thermostabilizing factors found in the proteins from thermophilic organisms has been performed since thermophilic proteins show substantially higher intrinsic thermostability than their counterparts from mesophilic organisms, while retaining the basic fold characteristic of the particular protein family [5, 6].

Although several comparative studies between thermophilic protein and their mesophilic counterparts suggested several thermostabilizing factors [7–11], the results revealed by comparative studies could not be accepted as general modes governing protein thermostability. The reason is that the results are quite various according to protein families, that is, each comparative study shows just case-by-case result in each protein family. To achieve more general patterns of protein thermostability, systematical analyses for investigating the common thermostabilizing factors highly found in several families of thermophilic proteins (not highly found in mesophilic ones) have been performed [12–17]. However, the systematic analyses have investigated so simple features, for example, total number of residual interactions or total size/volume of hydrophobic area or electronegative surface, etc., so that the findings were not helpful for understanding detailed thermostabilizing mechanisms, which could be used as practical guideline to engineer protein thermostability [1]. In other words, because the conventional systematical analyses were focused on too general features, it was difficult to find out the residue-level aspects or factors of protein thermostability.

In previous reports, we have tried to analyze the residue-level features of thermostabilizing factors in a systematical fashion by employing a concept of residual structure states estimated by packing value [18–20]. Considering residual structure states, we compared the distribution of amino acid types, residual properties, and structural features between thermophilic and mesophilic protein families. Finally, we found out distinctive residual patterns related with protein thermostability [18–20], that is, we showed how differently the amino acids, residual properties, and structural features are distributed in the thermophilic protein families from the mesophilic protein ones [18–20].

Herein, to find out the residual thermostabilizing factors, we approached it in a different way by investigating distinctive relative occurrences of amino acid kinds, residual properties, and secondary structure types in each residual structure state. We carried out a comparative analysis of 20 pairs of thermophilic and mesophilic proteins to achieve more explicit thermostabilizing factors. Through these analyses, we could get some clues on which residual factors in each residual structure state could contribute more to protein thermostability.

Material and Methods

Protein Model System

All the protein structural data used in this study were found in the Protein Data Bank (PDB) atomic coordinate database at the Research Collaboratory for Structural Bioinformatics

(RCSB) [21]. The criteria for selecting the sets of thermophilic and mesophilic protein structure pairs were based on our previous report [18–20], and the number of the sets for thermophilic and mesophilic protein pairs is 20. The list of the set used in this study is as follows (PDB code of thermophilic/mesophilic): 1zin/1aky, 1tmy/3chy, 1aj8/1csh, 1tfe/1efu_b, 1yna/1xnb, 1gtm/1hrd, 1hdg/1gad, 2prd/1ino, 1ldn/1ldg, 1bdm/4mdh, 3mds/1qmn, 1xgs/1mat, 3pfk/2pfk, 1php/1qpg, 1ebd/1lpf, 1ril/2rn2, 1caa/8rxn, 1thm/1st3, 1lnf/1npc, and 1btm/1ypi.

Residual Structure States

The residual structure states were determined by the residual packing value of each residue, which was calculated by an extension of the occluded surface algorithm [22, 23]. Residual packing value equals to 0.0 if there is no occluding van der Waals surface within 2.8 Å of the molecular surface; residual packing value equals to 1.0 if 100 % of the molecular surface were in contact with the van der Waals surface of other atoms. On the basis of the calculated residual packing values, five kinds of residual structure states were determined as follows: in the case of the fully exposed state (FE state), the residues have 0 to 0.15 residual packing value; in the exposed state (E state), the residues have 0.15 to 0.30; in the partially exposed (or partially buried) state (P state), the residues have 0.30 to 0.45; in the buried state (B state), the residues have 0.45 to 0.60; and in the well-buried state (WB state), the residues have 0.60 to 0.75.

Residual Properties Investigation

Flexible residues (Flex) and rigid residues (Rigid) were selected through comparing their α-carbon flexibility with one another. The residual α-carbon flexibility is calculated by obtaining the temperature B value of the α-carbon atoms in the PDB data [24]. In this study, flexible residues (Flex) were defined as the top 10 % residues showing the highest α-carbon flexibility among all the residues in a protein. Rigid residues (Rigid) were defined as the top 10 % residues showing the lowest α-carbon flexibility.

For describing the relationship of water solvation, the residues with high solvation energy (HSE) and low solvation energy (LSE) in the native state of protein were considered through comparing their residual solvation energy with one another. The residual solvation free energies calculated are based on the atomic solvation parameters of Eisenberg and McLachan [25]. For a given residue, the residual solvation free energy is determined as the sum of the contributions from the individual atoms (neglecting hydrogen atoms) in the residue taking into account their solvent-exposed surface area in the native state of protein. The residues with HSE were defined as the top 10 % residues showing the highest solvation energy among all the residues in a protein structure. The residues with LSE were defined as the top 10 % residues showing the lowest solvation energy.

The number of hydrogen bonds (Hbond) was calculated by counting the number of non-hydrogen atoms in each residue-involved donor and accepter hydrogen bonds, which are determined by the distance of donors and accepters within 4.0 Å [26]. Calculation of the number of hydrogen bonds was carried out by the portable module of biopolymer on SYBYL. In the case of salt bridges (SB), cation–pi interactions (Cat–pi), and disulfide bonds (SSbond), each number was estimated by the Protein Explorer package provided by the Department of Microbiology, University of Massachusetts Amherst. Salt bridges are assigned to two atoms of opposite charge, when the atoms were observed to be within 4.0 Å. Positively charged atoms include side chain N atom in ARG, LYS, and HIS, while

negatively charged atoms include side chain O atom in ASP and GLU [27]. Cation–pi interactions are assigned to aromatic residues, when a cationic side chain of ARG or LYS is near an aromatic side chain of PHE, TRP, or TYR. Ninety-nine percent of significant cation–pi interactions occur within a distance of 6.0 Å [28].

The secondary structure was determined by the Kabsch–Sander procedure [29]. The one-letter code of the secondary structure state was assigned to the residue as follows: E is the residue in the extended strand that participates in the beta ladder, B is in an isolated beta bridge, H is in the alpha helix, G is in the 3/10 helix, and T is in the hydrogen bonded turn. The residue irrelevant to any secondary structure (e.g., residues in loop) was not assigned with the one-letter code.

Statistical Analysis

The relative occurrence of residual factor in each residual structure state i was determined here as (occurrence of residual factor)/(total residue number of residual structure state i). To select a higher relative occurrence of residual factor in each residual structure state found in thermophilic proteins than in mesophilic proteins, test statistic for average relative occurrence (R_i) of both groups was carried out by calculation of t value. The t_i value is calculated as follows:

$$t_i = (R_{i-\mathrm{Th}} - R_{i-\mathrm{Me}})\big/\sqrt{\big(S^2_{i-\mathrm{Th}}/N_{\mathrm{Th}} + S^2_{i-\mathrm{Me}}/N_{\mathrm{Me}}\big)}, \tag{1}$$

where $R_{i\text{-Th}}$ and $R_{i\text{-Me}}$ are the average relative occurrences of residual factor in each residual structure state i of thermophilic protein groups and mesophilic protein groups, respectively; $S^2_{i\text{-Th}}$ and $S^2_{i\text{-Me}}$ are the deviations of average relative occurrences of R_i of thermophilic protein groups and mesophilic protein groups, respectively; and N_{Th} and N_{Me} are the total number of thermophilic protein groups and mesophilic protein groups, respectively. The degrees of freedom, df ($= N_{\mathrm{Th}} + N_{\mathrm{Me}} - 2$), is 38, in which the value is large enough to be considered as infinite sample sets. Critical levels of $t_{0.1}$ ($= 1.282$) and $t_{0.01}$ ($= 2.326$) in a one-tailed t test (with $df = $ infinite) were used as criteria for selecting the distinctive residual factor of thermophilic proteins [30].

Results and Discussion

Thermostabilizing Factors in Fully Exposed Residual State (FE State)

We investigated the relative occurrences of amino acids, residual properties, and secondary structure types in FE state and compared each statistic value between thermophilic and mesophilic protein groups as arranged in Table 1. t test results showed which residual factors in FE state play more important roles related to protein thermostability. Among the results, the relative occurrences of GLN, ILE, and PHE were observed to be distinctively higher in FE state of thermophilic proteins than those of mesophilic ones. t values of GLN, ILE, and PHE were investigated to be 1.4208, 1.6278, and 1.2833, respectively. The results belonging to other categories such as residual properties and secondary structure types showed no characteristic difference in FE state.

Compared to ASN and VAL in FE state, GLN with long alkyl group and ILE with long chain were observed to have more different relative occurrence in FE state of thermophilic proteins. In addition, PHE, an aromatic residue without polar group, showed high relative

Table 1 Different relative occurrences of amino acids, residual property, and secondary structure in fully exposed structure state (FE state) between thermophilic and mesophilic proteins

FE state	Thermo	Meso	t value
Amino acid			
ALA	0.0745±0.0291	0.0578±0.0032	0.5162
ARG	0.0669±0.0177	0.0495±0.0036	0.7296
ASN	0.0676±0.0199	0.0869±0.0058	−0.5901
ASP	0.1387±0.0503	0.1248±0.0110	0.1980
CYS	0.0000	0.0000	0.0000
GLN	0.0436±0.0134	0.0209±0.0019	**1.4208**
GLU	0.1042±0.0223	0.0954±0.0047	0.2875
GLY	0.1092±0.0287	0.0772±0.0045	0.9124
HIS	0.0251±0.0109	0.0166±0.0019	0.6145
ILE	0.0129±0.0079	0.0000	**1.6278**
LEU	0.0202±0.0095	0.0220±0.0140	−0.1052
LYS	0.1509±0.0401	0.1995±0.0315	−0.9531
MET	0.0179±0.0110	0.0181±0.0085	−0.0165
PHE	0.0271±0.0155	0.0063±0.0046	**1.2833**
PRO	0.0401±0.0134	0.0805±0.0338	−1.1091
SER	0.0461±0.0356	0.0708±0.0238	−0.5773
THR	0.0469±0.0248	0.0528±0.0138	−0.2074
TRP	0.0000	0.0022±0.0022	−0.9975
TYR	0.0032±0.0032	0.0045±0.0045	−0.2367
VAL	0.0049±0.0036	0.0141±0.0078	−1.0611
Residual property			
Flex	0.3737±0.0395	0.4385±0.0392	−1.1643
Rigid	0.0203±0.0115	0.0277±0.0121	−0.4424
HSE	0.2183±0.0369	0.2671±0.0400	−0.8969
LSE	0.3430±0.0518	0.3568±0.0442	−0.2030
Hbond	0.4619±0.0548	0.4710±0.0589	−0.1133
SB	0.0163±0.0089	0.0295±0.0108	−0.9463
Cat–pi	0.0000	0.0000	0.0000
SSbond	0.0000	0.0000	0.0000
Secondary structure			
E	0.0367±0.0182	0.0193±0.0080	0.8762
B	0.0000	0.0000	0.0000
H	0.1105±0.0256	0.0916±0.0219	0.5593
G	0.0774±0.0228	0.0800±0.0334	−0.0654
T	0.2941±0.0266	0.2955±0.0281	−0.0340

The bold t values are significant at the 10 % level [$t_{0.10}$ (= 1.282)]

Thermo the average relative occurrence of thermophilic protein groups (with standard deviation), *Meso* the average relative occurrence of mesophilic protein groups (with standard deviation)

occurrence in FE state compared with TRP and TYR, aromatic residues with polar group. Interestingly, GLN and ILE can be considered as the ones with higher flexibility and higher solvation energy than ASN and VAL. PHE can be also considered as the one with higher

solvation energy than TRP or TYR. Nevertheless, the residues with 10 % top highest flexibility (Flex) or the residues with 10 % top highest solvation energy (HSE) were investigated to have no distinctive differences. They were even observed to be less found in FE state of thermophilic proteins than those of mesophilic ones (−1.1643 and −0.8969 of t values, respectively). There are two explanations possible: One is that in fully exposed state, long-chained or more hydrophobic amino acids might play important roles related to protein thermostability among the amino acids with similar properties. However, they might not be the residues responsible for increasing the structural flexibility and solvation energy. The other is that thermophilic proteins may be more tolerant of such long-chained or more hydrophobic amino acids being placed in the FE state than mesophilic proteins.

Thermostabilizing Factors in Exposed Residual State (E State)

The relative occurrences of amino acids, residual properties, and secondary structure types were investigated in E state and statistically compared between thermophilic and mesophilic proteins as arranged in Table 2. In the case of amino acids, the relative occurrences of ALA, ARG, GLU, SER, and VAL in E state were observed to show distinctive values in thermophilic proteins compared to mesophilic ones. t values of ALA, ARG, GLU, SER, and VAL were investigated to be −2.3812, 1.8547, 1.8874, −1.3134, and −1.3300, respectively. In the case of residual properties, the relative occurrences of salt bridges (SB) and the residue with LSE were investigated to show characteristic differences related to protein thermostability (2.4225 and 1.5801 of t values, respectively). In the case of secondary structures, the relative occurrence of the residues in 3/10 helix (G) was observed to be highly found in thermophilic proteins (1.4897 of t value). In particular, the relative occurrences of salt bridges and ALA in E state showed big differences between thermophilic and mesophilic proteins since their t values are above the critical value of $t_{0.01}$ (= 2.326).

A significant increase in the number of salt bridges has been reported for most structures of thermostable proteins [9, 31, 32]. Our previous results also reported that salt bridge has higher preference for the exposed state of thermophilic proteins [18]. Salt bridges could be expected to stabilize the exposed part of a protein structure, which might be composed of more flexible and less stable structure than the buried part. Among the relative occurrences of salt bridge participants, only ARG and GLU showed a distinctive difference. Compared to other salt bridge participants (LYS or HIS and ASP), ARG and GLU could be expected to play important roles in exposed state of thermophilic proteins through constituting the salt bridge, which could be contributive to increasing the thermostability of the exposed structure. In addition, the charged residues were expected to contribute to thermostability since the residues with LSE were observed to have higher relative occurrence in the exposed state of thermophilic proteins than mesophilic ones. These results presented that in E state, the charged residues (in particular, ARG and GLU) could contribute to the electrostatic interactions and make the exposed interactions preferable for water molecules (that is, LSE), which can be energetically effective for stabilizing the local conformation of protein structure.

On the other hand, it is interesting that ALA, SER, and VAL were observed to have lower relative occurrences in E state of thermophilic proteins. The results that small aliphatic residues such as ALA and VAL were less found in E state of thermophilic proteins should be considered with the trend about higher occurrence of long-chained ILE or GLN in FE state of thermophilic proteins as mentioned in the above section, that is, in fully exposed and

Table 2 Different relative occurrences of amino acids, residual property, and secondary structure in exposed structure state (E state) between thermophilic and mesophilic proteins

E state	Thermo	Meso	t value
Amino acid			
ALA	0.0533±0.0085	0.0860±0.0024	***−2.3812***
ARG	0.0714±0.0110	0.0476±0.0015	**1.8547**
ASN	0.0577±0.0095	0.0774±0.0030	−1.2084
ASP	0.0777±0.0093	0.0781±0.0018	−0.0276
CYS	0.0023±0.0016	0.0014±0.0002	0.4848
GLN	0.0502±0.0080	0.0456±0.0016	0.4389
GLU	0.1504±0.0150	0.1128±0.0029	**1.8874**
GLY	0.1070±0.0094	0.1010±0.0019	0.4726
HIS	0.0234±0.0055	0.0190±0.0011	0.5916
ILE	0.0252±0.0063	0.0207±0.0010	0.5739
LEU	0.0333±0.0057	0.0270±0.0050	0.8324
LYS	0.1317±0.0113	0.1344±0.0143	−0.1473
MET	0.0146±0.0050	0.0123±0.0036	0.3621
PHE	0.0101±0.0027	0.0131±0.0034	−0.6949
PRO	0.0591±0.0087	0.0554±0.0083	0.3109
SER	0.0425±0.0111	0.0642±0.0122	**−1.3134**
THR	0.0468±0.0107	0.0533±0.0084	−0.4770
TRP	0.0022±0.0015	0.0015±0.0015	−0.2261
TYR	0.0170±0.0039	0.0152±0.0046	0.2979
VAL	0.0241±0.0043	0.0327±0.0049	**−1.3300**
Residual property			
Flex	0.2058±0.0138	0.1976±0.0097	0.4881
Rigid	0.0426±0.0084	0.0401±0.0074	0.2241
HSE	0.2061±0.0077	0.2018±0.0075	0.4021
LSE	0.2268±0.0096	0.2027±0.0119	**1.5801**
Hbond	1.2420±0.0312	1.1802±0.0445	1.1378
SB	0.1273±0.0151	0.0830±0.0104	***2.4225***
Cat–pi	0.0011±0.0011	0.0013±0.0012	−0.0675
SSbond	0.0000	0.0000	0.0000
Secondary structure			
E	0.0947±0.0224	0.1053±0.0196	−0.3560
B	0.0069±0.0033	0.0083±0.0028	−0.3144
H	0.3161±0.0380	0.3150±0.0386	0.0195
G	0.0726±0.0152	0.0438±0.0120	**1.4897**
T	0.2040±0.0172	0.2056±0.0170	−0.0662

The bold t values are significant at the 10 % level [$t_{0.10}$ (= 1.282)], and the bold italic t values are significant at 1 % [$t_{0.01}$ (= 2.326)]

Thermo the average relative occurrence of thermophilic protein groups (with standard deviation), *Meso* the average relative occurrence of mesophilic protein groups (with standard deviation)

exposed state of thermophilic proteins, the residues with shorter alkyl groups have low occurrence, whereas longer residues were observed to show the reverse. In the case of SER,

its molecular interaction, binding with surrounding waters, should be considered. Since the water, which could be interacted with SER (usually mediated by hydrogen bonding), would be released at higher temperature, the local protein structure around water-binding site such as SER could be changed and become unstable enough to evoke protein instability [33, 34].

In addition, the residues in G (3/10 helix) structures were investigated to show higher relative occurrences in E state of thermophilic proteins. Since 3/10 helix structure is one of the alpha-helix structure variants and this structure is expected to play an important role in the stabilization of the whole alpha-helix structure in the protein structure [35], such higher occurrence of 3/10 helix structure in E state of thermophilic proteins might be related to protein thermostability.

Thermostabilizing Factors in Partially Exposed (or Partially Buried) Residual State (P State)

We investigated the relative occurrences of amino acids, residual properties, and secondary structure types in P state and compared each statistic value between thermophilic and mesophilic protein groups as arranged in Table 3. t test was also performed to analyze which factors in P state play more important roles related to protein thermostability. In the case of amino acids, the relative occurrence of SER in P state was observed to have a distinctively lower value in thermophilic proteins (-2.6958 of t value). In the case of residual properties, the relative occurrence of the flexible residue (Flex) in P state was observed to be characteristic features of thermophilic proteins compared to mesophilic ones showing 2.7045 of t value. These two results could be considered as outstanding patterns related to protein thermostability since their t values are below the critical value of $-t_{0.01}$ ($= -2.326$) and above the critical value of $t_{0.01}$ ($= 2.326$), respectively.

In the case of SER, as mentioned in the above section, its binding interaction with surrounding waters may be involved. The water interacted with SER (usually mediated by hydrogen bonding) would affect protein instability when the water is released from the water-binding site at higher temperature [33, 34]. Although SER was investigated to have lower number in thermophilic proteins than in mesophilic ones, they showed the lowest relative occurrences especially in P state. Since the residues in P state have a higher chance to interact with neighboring residues and constitute the local protein structure among the residues in water-interacted parts (outer or boundary part), they also have higher chances to evoke protein instability in case that water is released at higher temperature.

In the case of local flexibility and rigidity of protein structure, several molecular interactions such as hydrophobic interaction, hydrogen bonding, salt bridge, or disulfide bond could be involved [36]. In P state, flexible residues have higher chances to interact with neighboring residues and to be stabilized by molecular interactions and forces. Therefore, this result indicated that the flexible residues were located more prevalently in P state of thermophilic proteins than mesophilic ones by such big differences.

Thermostabilizing Factors in Buried Residual State (B State)

The relative occurrences of amino acids, residual properties, and secondary structure types in B state were statistically analyzed between thermophilic and mesophilic proteins as arranged in Table 4. t test results showed which factors play more important roles in B state related to protein thermostability. In the case of amino acids, the relative occurrences of ARG, GLU, and MET were observed to show distinctive values in thermophilic proteins compared to

Table 3 Different relative occurrences of amino acids, residual property, and secondary structure in partially exposed (or partially buried) structure state (P state) between thermophilic and mesophilic proteins

P state	Thermo	Meso	t value
Amino acid			
ALA	0.0848±0.0083	0.0799±0.0018	0.4202
ARG	0.0573±0.0062	0.0503±0.0013	0.8123
ASN	0.0397±0.0056	0.0510±0.0016	−1.2281
ASP	0.0660±0.0055	0.0721±0.0017	−0.6407
CYS	0.0103±0.0062	0.0153±0.0017	−0.5098
GLN	0.0328±0.0051	0.0293±0.0010	0.5073
GLU	0.0681±0.0085	0.0633±0.0012	0.4790
GLY	0.0852±0.0062	0.0856±0.0014	−0.0481
HIS	0.0249±0.0048	0.0264±0.0010	−0.2407
ILE	0.0655±0.0073	0.0527±0.0016	1.2682
LEU	0.0678±0.0067	0.0741±0.0087	−0.5777
LYS	0.0720±0.0091	0.0623±0.0054	0.9128
MET	0.0251±0.0061	0.0207±0.0046	0.5784
PHE	0.0257±0.0043	0.0235±0.0031	0.4097
PRO	0.0485±0.0054	0.0440±0.0073	0.4884
SER	0.0389±0.0069	0.0624±0.0053	***−2.6958***
THR	0.0582±0.0061	0.0679±0.0100	−0.8289
TRP	0.0105±0.0031	0.0039±0.0039	−0.8927
TYR	0.0446±0.0080	0.0359±0.0064	0.8543
VAL	0.0741±0.0066	0.0685±0.0044	0.7008
Residual property			
Flex	0.0912±0.0064	0.0651±0.0072	***2.7045***
Rigid	0.0867±0.0109	0.0777±0.0075	0.6774
HSE	0.1231±0.0078	0.1162±0.0065	0.6737
LSE	0.0912±0.0085	0.0886±0.0079	0.2308
Hbond	1.6797±0.0468	1.7088±0.0398	−0.4728
SB	0.1180±0.0110	0.1019±0.0106	1.0555
Cat–pi	0.0070±0.0028	0.0103±0.0042	−0.6525
SSbond	0.0011±0.0008	0.0003±0.0003	0.9235
Secondary structure			
E	0.1905±0.0287	0.1790±0.0292	0.2804
B	0.0119±0.0045	0.0218±0.0114	−0.8095
H	0.3294±0.0361	0.3522±0.0344	−0.4577
G	0.0376±0.0077	0.0307±0.0058	0.7160
T	0.0971±0.0107	0.1034±0.0058	−0.5164

The bold italic t values are significant at 1 % [$t_{0.01}$ (= 2.326)]

Thermo the average relative occurrence of thermophilic protein groups (with standard deviation), *Meso* the average relative occurrence of mesophilic protein groups (with standard deviation)

mesophilic ones. t values of ARG, GLU, and MET were investigated to be 1.3206, 1.9562, and −1.6770, respectively. Other results belonging to residual properties and secondary structure types showed no characteristic occurrences in B state.

Table 4 Different relative occurrences of amino acids, residual property, and secondary structure in buried structure state (B state) between thermophilic and mesophilic proteins

B state	Thermo	Meso	t value
Amino acid			
ALA	0.1283±0.0121	0.1145±0.0023	0.8722
ARG	0.0153±0.0023	0.0114±0.0004	**1.3206**
ASN	0.0253±0.0036	0.0270±0.0014	−0.2331
ASP	0.0343±0.0051	0.0290±0.0012	0.7306
CYS	0.0130±0.0031	0.0128±0.0008	0.0487
GLN	0.0129±0.0025	0.0153±0.0007	−0.5897
GLU	0.0312±0.0058	0.0181±0.0008	**1.9562**
GLY	0.0804±0.0074	0.0823±0.0013	−0.1956
HIS	0.0141±0.0039	0.0169±0.0006	−0.5817
ILE	0.1146±0.0102	0.1000±0.0025	0.9747
LEU	0.1248±0.0150	0.1321±0.0130	−0.3677
LYS	0.0086±0.0029	0.0101±0.0024	−0.4148
MET	0.0198±0.0032	0.0291±0.0045	**−1.6770**
PHE	0.0481±0.0049	0.0584±0.0077	−1.1291
PRO	0.0265±0.0039	0.0308±0.0045	−0.7178
SER	0.0339±0.0055	0.0355±0.0053	−0.2071
THR	0.0415±0.0049	0.0437±0.0050	−0.3154
TRP	0.0314±0.0104	0.0077±0.0077	0.4383
TYR	0.0595±0.0104	0.0557±0.0126	0.2299
VAL	0.1364±0.0098	0.1515±0.0106	−1.0438
Residual property			
Flex	0.0199±0.0057	0.0216±0.0046	−0.2360
Rigid	0.1604±0.0102	0.1788±0.0145	−1.0417
HSE	0.0038±0.0016	0.0016±0.0009	1.2310
LSE	0.0070±0.0023	0.0117±0.0030	−1.2550
Hbond	1.9345±0.0305	1.8967±0.0334	0.8379
SB	0.0413±0.0061	0.0400±0.0051	0.1696
Cat–pi	0.0155±0.0055	0.0150±0.0045	0.0666
SSbond	0.0011±0.0008	0.0003±0.0003	0.8462
Secondary structure			
E	0.3251±0.0391	0.3525±0.0360	−0.5152
B	0.0160±0.0089	0.0153±0.0044	0.0708
H	0.3660±0.0427	0.3903±0.0397	−0.4172
G	0.0250±0.0058	0.0244±0.0059	0.0731
T	0.0393±0.0070	0.0357±0.0054	0.4097

The bold t values are significant at the 10 % level [$t_{0.10}$ (= 1.282)]

Thermo the average relative occurrence of thermophilic protein groups (with standard deviation), *Meso* the average relative occurrence of mesophilic protein groups (with standard deviation)

The results of ARG and GLU were expected to get involved in their contribution to electrostatic interaction in the inner part of the protein structure. Although the relative occurrences of hydrogen bond and salt bridge in B state were observed to have no difference

between thermophilic and mesophilic proteins, ARG and GLU might play prevalent roles in making electrostatic interaction in thermophilic proteins since the salt bridges mediated by ARG–GLU paring were investigated to be effective for stabilizing protein structure. In the case of MET, it can be expected that in B state, MET would be less found in thermophilic proteins than in mesophilic proteins since its sulfur group could influence the internal network of hydrophobic interactions, although MET showed hydrophobic property.

Thermostabilizing Factors in Well-Buried Residual State (WB State)

We investigated the relative occurrences of amino acids, residual properties, and secondary structure types in WB state and compared each statistic values between thermophilic and mesophilic protein groups as arranged in Table 5. In the case of amino acids, the relative occurrences of ALA, ASP, and GLY were observed to show distinctive values in thermophilic proteins compared to mesophilic proteins. t values of ALA, ASP, and GLY were investigated to be 1.3967, −1.3908, and −1.7916, respectively. In the case of residual properties, the relative occurrence of cation–pi interaction (Cat–pi) was observed to show characteristic value in WB state of thermophilic proteins showing 1.3708 of t value. In the case of secondary structures, the relative occurrences of the extended beta strand (E) and 3/10 helix (G) were observed to show distinctive structural difference between thermophilic and mesophilic proteins. t values of E and G were investigated to be −1.3511 and 1.4529, respectively.

GLY is known as the residue for making void volume or cavity in the inner part of the protein structure [34]. However, most of the other systematical analyses have reported that there is no typical pattern of GLY occurrence in thermophilic proteins. In our other study, GLY was observed to show no structural difference in distribution between thermophilic and mesophilic proteins [19]. However, this study showed that in WB state, GLY was less found in thermophilic proteins than mesophilic ones. Since GLY has a higher chance to make void volume or cavity in the inner part of the protein structure, GLY was expected to be less located in thermophilic proteins. Compared to GLY, ALA showed higher occurrence in WB state of thermophilic proteins than that of mesophilic ones. It can be expected that ALA would be more preferable to contribute to the increased hydrophobic packing than GLY with no methyl group. In the case of ASP, small-charged amino acid such as ASP was less found in WB state of thermophilic protein than that of mesophilic ones compared to large-charged amino acid such as GLU.

Cation–pi interaction is assigned when aromatic side chain of PHE, TRP, and TYR is near a cationic side chain of ARG and LYS [28]. Like salt bridges, cation–pi interactions are also subject to their location and geometry in protein structure. Among the aromatic amino acids, TRP is the most likely to be involved in a cation–pi interaction. Higher relative occurrence of TRP in WB of thermophilic proteins might be correlated with higher relative occurrence of cation–pi interaction.

In the case of secondary structure types, the extended beta strand (E) was less found in WB state of thermophilic proteins. This result might be related with the fact that the residue might have difficulties in keeping the extended beta strand in the dense packing structure of the inner part of the protein [35]. On the other hand, the residues in 3/10 helix (G) were investigated to show higher relative occurrences in WB state of thermophilic proteins. As mentioned in the previous section, 3/10 helix, one of the alpha-helix structure variants, can be expected to contribute to stabilizing the whole alpha-helix structure [35], that is, the high occurrence of the residues in 3/10 helix in WB state together with those in E state would be related to protein thermostability.

Table 5 Different relative occurrences of amino acids, residual property, and secondary structure in well-buried structure state (WB state) between thermophilic and mesophilic proteins

WB state	Thermo	Meso	*t* value
Amino acid			
ALA	0.2968±0.0454	0.2083±0.0099	**1.3967**
ARG	0.0000	0.0000	0.0000
ASN	0.0118±0.0071	0.0033±0.0007	1.0864
ASP	0.0089±0.0059	0.0461±0.0058	**−1.3908**
CYS	0.0705±0.0359	0.0854±0.0115	−0.2368
GLN	0.0071±0.0071	0.0000	1.0002
GLU	0.0139±0.0085	0.0074±0.0011	0.6567
GLY	0.2027±0.0371	0.3255±0.0129	**−1.7916**
HIS	0.0257±0.0132	0.0196±0.0032	0.3129
ILE	0.0116±0.0065	0.0196±0.0025	−0.6248
LEU	0.0308±0.0140	0.0375±0.0187	−0.2835
LYS	0.0063±0.0062	0.0000	1.0002
MET	0.0142±0.0083	0.0146±0.0101	−0.0320
PHE	0.0241±0.0135	0.0295±0.0144	−0.2743
PRO	0.0425±0.0256	0.0096±0.0054	1.2560
SER	0.0564±0.0155	0.0916±0.0227	−1.2815
THR	0.0235±0.0102	0.0339±0.0159	−0.5532
TRP	0.0142±0.0125	0.0000	1.1312
TYR	0.0219±0.0096	0.0118±0.0068	0.8566
VAL	0.0671±0.0181	0.0562±0.0187	0.4170
Residual property			
Flex	0.0114±0.0079	0.0440±0.0258	−1.2085
Rigid	0.2009±0.0380	0.2727±0.0693	−0.9084
HSE	0.0000	0.0000	0.0000
LSE	0.0000	0.0000	0.0000
Hbond	1.8643±0.1169	2.0096±0.0778	−1.0347
SB	0.0096±0.0047	0.0315±0.0251	−0.8580
Cat–pi	0.0188±0.0137	0.0000	**1.3708**
SSbond	0.0000	0.0000	0.0000
Secondary structure			
E	0.1920±0.0477	0.3023±0.0662	**−1.3511**
B	0.0134±0.0074	0.0100±0.0055	0.3740
H	0.5577±0.0654	0.5765±0.0707	−0.1950
G	0.0188±0.0129	0.0000	**1.4529**
T	0.0274±0.0112	0.0251±0.0117	0.1413

The bold *t* values are significant at the 10 % level [$t_{0.10}$ (= 1.282)]

Thermo the average relative occurrence of thermophilic protein groups (with standard deviation), *Meso* the average relative occurrence of mesophilic protein groups (with standard deviation)
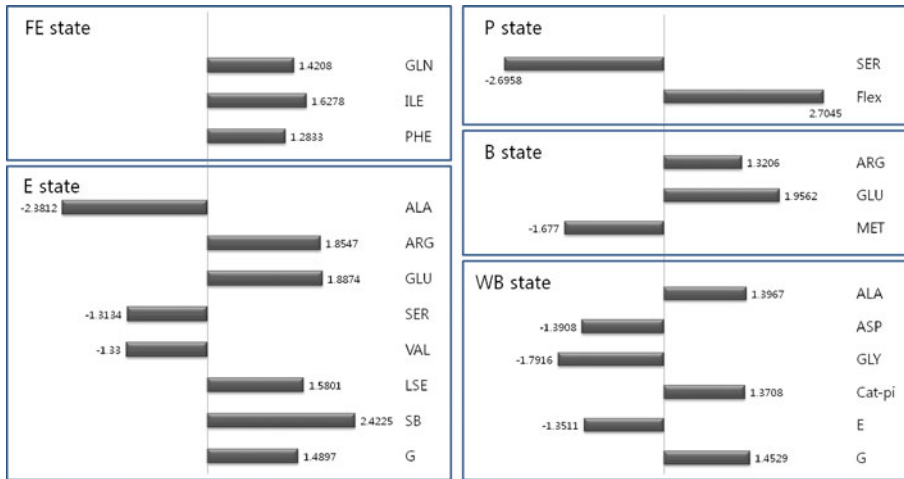
Fig. 1 Summary of *t* values for the thermostabilizing factors distinctively found in thermophilic proteins

Conclusion

This study analyzed the different relative occurrences of amino acids, residual properties, and secondary structure types between thermophilic and mesophilic proteins according to the residual structure states such as fully exposed, exposed, partially exposed (or partially buried), buried, and well-buried state. Through statistical analysis, several important factors related to protein thermostability were suggested in each residual structure state as follows (also as arranged in Fig. 1): (1) in the case of fully exposed state, higher relative occurrences of GLN, ILE, and PHE; (2) in the case of exposed state, higher relative occurrences of ARG, GLU, salt bridges, the residue with LSE, and the residues in 3/10 helix, and lower relative occurrences of ALA, SER, and VAL; (3) in the case of partially exposed (or partially buried) state, higher relative occurrence of flexible residue and lower relative occurrence of SER; (4) in the case of buried state, higher relative occurrences of ARG and GLU, and lower relative occurrence of MET; and (5) in the case of well-buried state, higher relative occurrences of ALA, cation–pi interaction, the residues in 3/10 helix, and lower relative occurrences of GLY, ASP, and the residues in the extended beta strand. Compared to other systematical analyses, this study showed several factors distinctively found in each residual structural state, which could be considered as significant modes related to protein thermostability. These results could be used as guidelines to understand the structural basis of thermophilic protein structure and develop rational design strategies for protein thermostabilization.

References

1. van den Burg, B., & Eijsink, V. G. (2002). Selection of mutations for increased protein stability. *Current Opinion in Biotechnology, 13*, 333–337.
2. Fagain, C. O. (1995). Understanding and increasing protein stability. *Biochimica et Biophysica Acta, 1252*, 1–14.
3. Gromiha, M. M., Oobatake, M., Kono, H., Uedaira, H., & Sarai, A. (1999). Relationship between amino acid properties and protein stability: buried mutations. *Journal of Protein Chemistry, 18*, 565–578.
4. Dahiyat, B. I. (1999). In silico design for protein stabilization. *Current Opinion in Biotechnology, 10*, 387–390.
5. Jaenicke, R. (1996). Stability and folding of ultrastable proteins: eye lens crystallins and enzymes from thermophiles. *The FASEB Journal, 10*, 84–92.
6. Scandurra, R., Consalvi, V., Chiaraluce, R., Politi, L., & Engel, P. C. (1998). Protein thermostability in extremophiles. *Biochimie, 80*, 933–941.
7. Hecht, K., Wrba, A., & Jaenicke, R. (1989). Catalytic properties of thermophilic lactate dehydrogenase and halophilic malate dehydrogenase at high temperature and low water activity. *European Journal of Biochemistry, 183*, 69–74.
8. Russell, R. J., Hough, D. W., Danson, M. J., & Taylor, G. L. (1994). The crystal structure of citrate synthase from the thermophilic archaeon, *Thermoplasma acidophilum*. *Structure, 2*, 1157–1167.
9. Yip, K. S., Stillman, T. J., Britton, K. L., Artymiuk, P. J., Baker, P. J., Sedelnikova, S. E., et al. (1995). The structure of *Pyrococcus furiosus* glutamate dehydrogenase reveals a key role for ion-pair networks in maintaining enzyme stability at extreme temperatures. *Structure, 3*, 1147–1158.
10. Tanner, J. J., Hecht, R. M., & Krause, K. L. (1996). Determinants of enzyme thermostability observed in the molecular structure of *Thermus aquaticus* D-glyceraldehyde-3-phosphate dehydrogenase at 25 Angstroms resolution. *Biochemistry, 35*, 2597–2609.
11. Shiraki, K., Nishikori, S., Fujiwara, S., Hashimoto, H., Kai, Y., Takagi, M., et al. (2001). Comparative analyses of the conformational stability of a hyperthermophilic protein and its mesophilic counterpart. *European Journal of Biochemistry, 268*, 4144–4150.
12. Spassov, V. Z., Karshikoff, A. D., & Ladenstein, R. (1995). The optimization of protein-solvent interactions: thermostability and the role of hydrophobic and electrostatic interactions. *Protein Science, 4*, 1516–1527.
13. Vogt, G., & Argos, P. (1997). Protein thermal stability: hydrogen bonds or internal packing? *Folding and Design, 2*, S40–S46.
14. Vogt, G., Woell, S., & Argos, P. (1997). Protein thermal stability, hydrogen bonds, and ion pairs. *Journal of Molecular Biology, 269*, 631–643.
15. Karshikoff, A., & Ladenstein, R. (1998). Proteins from thermophilic and mesophilic organisms essentially do not differ in packing. *Protein Engineering, 11*, 867–872.
16. Chakravarty, S., & Varadarajan, R. (2000). Elucidation of determinants of protein stability through genome sequence analysis. *FEBS Letters, 470*, 65–69.
17. Szilagyi, A., & Zavodszky, P. (2000). Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits: results of a comprehensive survey. *Structure, 8*, 493–504.
18. Pack, S. P., & Yoo, Y. J. (2003). Protein thermostability: structure-based difference of residual properties between thermophilic and mesophilic proteins. *Journal of Molecular Catalysis B: Enzymatic, 26*, 257–264.
19. Pack, S. P., & Yoo, Y. J. (2004). Protein thermostability: structure-based difference of amino acid between thermophilic and mesophilic proteins. *Journal of Biotechnology, 111*, 269–277.
20. Pack, S. P., & Yoo, Y. J. (2005). Packing-based difference of structural features between thermophilic and mesophilic proteins. *International Journal of Biological Macromolecules, 35*, 169–174.
21. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., et al. (2000). The Protein Data Bank. *Nucleic Acids Research, 28*, 235–242.
22. Pattabiraman, N., Ward, K. B., & Fleming, P. J. (1995). Occluded molecular surface: analysis of protein packing. *Journal of Molecular Recognition, 8*, 334–344.
23. Fleming, P. J., & Richards, F. M. (2000). Protein packing: dependence on protein size, secondary structure and amino acid composition. *Journal of Molecular Biology, 299*, 487–498.
24. Parthasarathy, S., & Murthy, M. R. N. (1997). Analysis of temperature factor distribution in high-resolution protein structures. *Protein Science, 6*, 2561–2567.
25. Eisenberg, D., & McLachlan, A. D. (1986). Solvation energy in protein folding and binding. *Nature, 319*, 199–203.
26. McDonald, I. K., & Thornton, J. M. (1994). Satisfying hydrogen-bonding potential in proteins. *Journal of Molecular Biology, 238*, 777–793.

27. Dill, K. A. (1990). Dominant forces in protein folding. *Biochemistry, 29*, 7133–7155.
28. Gallivan, J. P., & Dougherty, D. A. (1999). Cation-pi interactions in structural biology. *Proceedings of the National Academy of Sciences of the United States of America, 96*, 9459–9464.
29. Kabsch, W., & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers, 22*, 2577–2637.
30. Mendenhall, W., & Beaver, R. J. (1991). *Introduction to probability and statistics*. Boston: PWS-Kent.
31. Korndorfer, I., Steipe, B., Huber, R., Tomschy, A., & Jaenicke, R. (1995). The crystal structure of holo-glyceraldehyde-3-phosphate dehydrogenase from the hyperthermophilic bacterium *Thermotoga maritima* at 2.5 A resolution. *Journal of Molecular Biology, 246*, 511–521.
32. Kelly, C. A., Nishiyama, M., Ohnishi, Y., Beppu, T., & Birktoft, J. J. (1993). Determinants of protein thermostability observed in the 1.9-A crystal structure of malate dehydrogenase from the thermophilic bacterium *Thermus flavus. Biochemistry, 32*, 3913–3922.
33. Denisov, V. P., Venu, K., Peters, J., Horlein, H. D., & Halle, B. (1997). Orientational disorder and entropy of water in protein cavities. *The Journal of Physical Chemistry. B, 101*, 9380–9389.
34. Nagendra, H. G., Sukumar, N., & Vijayan, M. (1998). Role of water in plasticity, stability, and action of proteins: the crystal structures of lysozyme at very low levels of hydration. *Proteins, 32*, 229–240.
35. Creighton, T. E. (1997). *Proteins: structures and molecular properties* (2nd ed.). New York: Freeman & Company.
36. Fields, P. A. (2001). Review: protein function at thermal extremes: balancing stability and flexibility. *Comparative Biochemistry and Physiology. Part A, 129*, 417–431.