Comparative analysis of protein thermostability: Differences in amino acid content and substitution at the surfaces and in the core regions of thermophilic and mesophilic proteins

Kiyonobu Yokota, Kenji Satou, Shin-ya Ohki *

Japan Advanced Institute of Science and Technology, Centre for Nano Materials and Technology, 1-1 Asahidai, Tatsunokuchi, Ishikawa 923-1292, Japan

Received 31 October 2005; received in revised form 30 March 2006; accepted 31 March 2006

Abstract

In order to investigate the factors responsible for protein thermostability, we performed a comparative analysis. For this study, we prepared a new dataset composed of 47 homologous pairs of thermophilic and mesophilic proteins. It is the largest comparative study dataset ever presented. The frequency and substitution preference of each amino acid type in the dataset were analyzed. Two kinds of residual structural states were considered, i.e. surface (solvent-exposed) and core (buried) regions. On the surface of thermophilic proteins, higher frequencies were observed for Arg, Glu, and Tyr. Analysis of substitution preference also suggests that these often appear by replacement of other amino acid types. The results indicate that Arg, Glu, and Tyr are suitable for location on the surface of thermophilic proteins. On the other hand, at the core of thermophilic proteins, Ala is often appeared. In addition, our t-test analysis provides the first quantitative information about trends in the frequencies and substitution preferences for Cys, Gln, Met, and Ser. The results indicate that Gln and Met on the surface and Cys and Ser in the core are disadvantageous for protein thermostability.

q 2006 NIMS and Elsevier Ltd. All rights reserved.

Keywords: Protein; Thermostability; Comparative analysis; Substitution preference

1. Introduction

Biomolecules are well organized materials whose size is on the nanoscale. They are very attractive targets for utilization as components of manufactured products. For example, they are expected to be the elements of sensors, switches, and so on. For such purposes, thermostability of the proteins would be a useful property. Therefore, researchers in both scientific and industrial fields are interested in the factors responsible for protein thermostability. Various means for control of the thermostability have been studied and these include amino acid charge [1,2], electrostatic interactions [3–8], aromatic interactions [9], compactness [10], deamidation [11], and so on.

To determine the factors responsible for protein thermostability, both experimental and theoretical methods have been employed. In the experimental approaches, proteins classified in the same protein family but living at different temperatures, have been compared [6,12]. The differences in thermostability observed by some physicochemical methods have often been discussed based on the amino acid substitution. Comparative studies have also been attempted using theoretical approaches [10,13–17]. Statistical investigations have been applied to reveal differences in the contents of amino acid types between thermophilic and mesophilic proteins. After such extensive studies, the mechanism of protein thermostability at atomic resolution has been gradually clarified. Although, the development of faster computers and user-friendly software packages has supported some of the research in bioinformatics, there remained the considerable problem that the number of structural data available was too small to construct a reliable dataset. However, this problem has been solved by recent substantial progress in structural genomics.

For our comparative study, we constructed a new dataset using the updated databases. In our dataset, there are 47 homologous pairs of thermophilic and mesophilic proteins. The 47 categories represent 47 different types of protein folds, and each protein pair was selected based on their sequence identity. The number of categories in our data set, at 47, is double that of previous similar studies [10,14,16,17]. For each amino acid type, the content and the substitution preference from mesophilic to thermophilic proteins were analyzed. In this
study, we considered two kinds of structural locations of amino acid residues at the surface and core regions in protein structures. The new findings using the large dataset are presented in this paper.

2. Material and methods

2.1. Dataset and definition of location of residues

To create the new dataset, we used three kinds of databases: the protein data bank (PDB) to obtain protein coordinates [18], the structural classification of proteins (SCOP) to identify the protein folding [19], and microorganisms and cell cultures (DSMZ) (http://www.dsmz.de/index.htm) to check the growth temperature of the source organisms.

Before assembling the dataset, we excluded some of the protein coordinates in order to avoid artificial bias. The criteria used were as follows: (i) relatively small proteins (<100 amino acid residues) were excluded to avoid their size effect, (ii) PDB entries containing ‘MUTATION’ in ‘COMPND’ fields were also excluded to avoid the mutation effect, (iii) low-resolution structures (>2.5 Å) were omitted to obtain precise data, and (iv) all-alpha and all-beta protein structures were excluded to eliminate fold redundancy, since these folds can often be found in certain regions of alpha/beta and alpha+beta proteins.

A dataset of homologous thermophilic and mesophilic proteins was constructed using the following procedure. First, we defined a protein with a growth temperature written in the DSMZ ‘information’ field of more than 50°C as thermophilic. For this categorization, a direct relationship, \( T_m = 24.4 + 0.93 T_{en} \), was used [16], where \( T_m \) is the melting temperature of a protein, and \( T_{en} \) is the average environmental temperature of the source organisms. Second, all thermophilic proteins were extracted from SCOP using the keyword ‘thermophilic organisms’. Third, all amino acid sequences of the thermophilic proteins selected in the preceding steps were checked for redundancy. Fourth, mesophilic counterparts of the thermophilic proteins remaining in the previous step were found using BLAST [20]. BLAST is a program for identification of protein homologues. Fifth, if more than one amino acid sequence hit by the BLAST search was classified as an identical protein fold, we employed a pair of thermophilic and mesophilic proteins with the best match in their amino acid sequences. This step was performed to avoid the bias of an evolutorial effect. Thus, in our dataset, onefold category has one pair of thermophilic and mesophilic proteins. Finally, we constructed the dataset containing 47 kinds of thermophilic and mesophilic protein pairs. Additionally, we defined the location of the residues in the protein structures. The location of a residue with a more than 5% relative solvent accessibility was defined as the solvent-exposed (surface). When a residue has a less than 5% relative solvent accessibility, the location was defined as the core (buried).

2.2. Statistical analysis

We carried out \( t \)-tests to estimate quantitatively the difference in amino acid content between thermophilic and mesophilic proteins. The \( t \)-test parameter, \( t_i \), can be described using Eq. (1).

\[
t_i = \frac{(X_T^i - X_M^i)}{\sqrt{(S_{T^2}^i/N_T) + (S_{M^2}^i/N_M)}}
\]

In (1), \( X_T^i \) and \( X_M^i \) are the frequency of amino acids in a certain region, \( i \) (i.e. surface or core), of the thermophilic and mesophilic protein groups, respectively, \( S_{T^2}^i \) and \( S_{M^2}^i \) are the deviations of frequency \( X_i \) of the thermophilic and mesophilic protein groups, respectively, and \( N_T \) and \( N_M \) are the total number of thermophilic and mesophilic protein groups, respectively.

The value \( t_i \) can be used to judge whether there is a statistical difference in the \( X_i \) values between thermophilic and mesophilic proteins. The \( t_i \) parameter is considered to be a reliability indicator.

2.3. Substitution preference in amino acid sequences for thermostability

To find the difference in amino acid sequences between thermophilic and mesophilic proteins, we aligned the amino acid sequences of each protein pair in the dataset. The alignment was performed using the ALIGN program [21]. Subsequently, the number of possible substitutions was calculated for each amino acid type. The values of substitution preference were normalized with the total number of mutation points.

3. Results and discussion

3.1. The protein dataset

Three kinds of databases were used to construct our new dataset. The protein coordinates were obtained from the PDB released on July 1, 2005, which contains 31,535 protein structures. The protein folds were obtained from SCOP version 1.67, released on February 2005, which contains 887 folds. The environment temperatures for the proteins were obtained from the DSMZ released on July 2002. We constructed the thermophilic–mesophilic protein dataset according to the procedure described above. Our dataset has 47 different types of protein folds, thus containing 94 protein structures. The contents of our dataset are shown in Table 1. The number of categories, at 47, is much greater than any of the other previous comparative studies [17]. Hence, we believe that our present dataset should provide more reliable results than those obtained in previous studies.

Differences in the amino acid contents between mesophilic and thermophilic proteins in the dataset were calculated. In this calculation, two kinds of residual structural states such as surface and core in protein structures were considered. The analysis was performed using Eq. (1) and the results are summarized in Table 2.

We also calculated the amino acid substitution preferences from mesophilic to thermophilic proteins. The results are summarized in Tables 3 and 4. Eight thousand and one hundred and eighty possible substitutions were detected, which is an extremely large number compared to a previous systematical study [16]. Larger values appearing in Tables 3 and 4 mean that the
substitution is favored when the mutations from mesophilic to thermophilic proteins are considered. It is thought that conversions from various amino acid types to a particular one indicate thermal adaptation by that substitution. In contrast, when substitutions of a specific amino acid by a variety of amino acids are observed, the starting amino acid type is judged unfavorable for thermostability.

3.2. Protein surface amino acids that favor thermostability

As shown in Table 2, the t-values for Arg, Glu, and Tyr at the protein surface are remarkably high (3.37, 4.04, and 2.03, respectively). This indicates clearly that these amino acid residues prefer to reside at the surface of thermophilic proteins.

It is believed that location of Arg at the protein surface is sufficient for thermostability. Arg forms ion pair or ion pair networks in protein structures [22,23], and it is this property that is related to protein thermostability [10,14,17,24,25]. The t-value of solvent-exposed Arg quantitatively supports this idea. Furthermore, as shown in Table 3, it was found that the 16 amino acid types at the protein surface show a preference for conversion to Arg. In particular, Ala and Glu at the protein surface, whose types at the protein surface show a preference for conversion to solvent-exposed Arg quantitatively supports this idea. Further-
|            | Entire          | Core          | Surface         | Entire          | Core          | Surface         |
|------------|-----------------|---------------|-----------------|-----------------|---------------|-----------------|
|            | a.a. | Mean | SD | Mean | SD | Mean | SD | Mean | SD | Mean | SD | Mean | SD |
| THERMO G   | 8.08 | 1.87 | 2.12 | 1.14 | 5.96 | 1.56 | K   | 6.62 | 2.83 | 0.16 | 0.22 | 6.46 | 2.82 |
| Meso       | 8.37 | 2.11 | 2.23 | 1.14 | 6.13 | 1.61 | 6.03 | 2.26 | 0.16 | 0.23 | 5.87 | 2.28 |
| T-value    | -1.10 | -0.87 | -0.82 | 1.25 | 1.25 | 1.25 | 1.25 | 1.25 | 1.25 | 1.25 | 1.25 | 1.25 |
| THERMO A   | 9.37 | 3.30 | 4.81 | 2.25 | 4.56 | 2.54 | R   | 5.75 | 2.41 | 0.17 | 0.24 | 5.58 | 2.44 |
| Meso       | 9.11 | 2.64 | 4.29 | 1.82 | 5.87 | 2.27 | 4.61 | 2.05 | 0.14 | 0.23 | 4.47 | 2.03 |
| T-value    | -1.56 | 2.81 | -3.32 | 3.86 | 1.08 | 3.37 | 1.08 | 3.37 | 1.08 | 3.37 | 1.08 | 3.37 |
| THERMO V   | 8.36 | 1.99 | 4.55 | 1.65 | 3.17 | 1.47 | H   | 5.07 | 2.01 | 0.40 | 0.41 | 5.66 | 2.01 |
| Meso       | 10.11 | 2.64 | 4.29 | 1.82 | 5.87 | 2.27 | 4.61 | 2.05 | 0.14 | 0.23 | 4.47 | 2.03 |
| T-value    | -1.20 | -0.82 | -3.32 | 3.86 | 1.08 | 3.37 | 1.08 | 3.37 | 1.08 | 3.37 | 1.08 | 3.37 |
| THERMO L   | 9.29 | 2.97 | 4.27 | 1.79 | 5.02 | 2.07 | D   | 5.44 | 1.62 | 0.62 | 0.54 | 4.81 | 1.62 |
| Meso       | 8.69 | 2.21 | 4.02 | 1.39 | 4.67 | 1.51 | 5.59 | 1.56 | 0.55 | 0.51 | 5.15 | 1.62 |
| T-value    | 1.48 | 1.02 | -0.79 | 1.27 | 1.27 | 1.27 | 1.27 | 1.27 | 1.27 | 1.27 | 1.27 | 1.27 |
| THERMO I   | 6.41 | 2.94 | 3.37 | 1.58 | 3.04 | 1.71 | E   | 8.08 | 2.55 | 0.41 | 0.54 | 7.67 | 2.63 |
| Meso       | 6.51 | 1.86 | 3.53 | 1.23 | 2.97 | 1.29 | 6.60 | 1.60 | 0.41 | 0.49 | 6.19 | 1.63 |
| T-value    | -0.23 | -0.72 | -0.72 | 4.08 | 0.14 | 4.04 | 0.14 | 4.04 | 0.14 | 4.04 | 0.14 | 4.04 |
| THERMO M   | 1.82 | 1.25 | 0.80 | 0.67 | 1.02 | 0.97 | S   | 4.43 | 1.88 | 0.88 | 0.67 | 3.55 | 1.66 |
| Meso       | 2.30 | 1.32 | 0.91 | 0.69 | 1.39 | 1.06 | 5.32 | 2.22 | 1.18 | 0.88 | 4.14 | 1.97 |
| T-value    | 2.39 | 1.02 | -2.49 | -2.49 | 1.08 | 1.08 | 1.08 | 1.08 | 1.08 | 1.08 | 1.08 | 1.08 |
| THERMO P   | 4.60 | 1.77 | 0.77 | 0.62 | 3.83 | 1.67 | T   | 5.02 | 1.70 | 1.46 | 0.93 | 3.56 | 1.63 |
| Meso       | 4.35 | 1.37 | 0.76 | 0.57 | 3.59 | 1.17 | 5.41 | 1.48 | 1.43 | 0.83 | 3.98 | 1.38 |
| T-value    | 1.08 | 0.11 | 1.10 | 1.49 | 0.24 | 1.50 | 0.24 | 1.50 | 0.24 | 1.50 | 0.24 | 1.50 |
| THERMO F   | 3.79 | 1.42 | 1.56 | 0.93 | 2.23 | 1.10 | N   | 3.43 | 1.95 | 0.38 | 0.42 | 3.05 | 1.85 |
| Meso       | 3.56 | 1.50 | 1.44 | 0.84 | 2.13 | 1.22 | 3.79 | 1.57 | 0.40 | 0.46 | 3.38 | 1.48 |
| T-value    | 0.89 | 0.98 | 0.50 | 1.26 | 0.47 | 1.25 | 0.47 | 1.25 | 0.47 | 1.25 | 0.47 | 1.25 |
| THERMO Y   | 3.22 | 1.65 | 0.66 | 0.69 | 2.56 | 1.22 | Q   | 2.64 | 1.64 | 0.24 | 0.36 | 2.40 | 1.49 |
| Meso       | 2.94 | 1.35 | 0.71 | 0.62 | 2.24 | 1.08 | 3.62 | 1.49 | 0.22 | 0.37 | 3.41 | 1.43 |
| T-value    | 1.52 | 0.54 | 2.83 | 2.88 | 0.75 | 3.05 | 0.75 | 3.05 | 0.75 | 3.05 | 0.75 | 3.05 |
| THERMO W   | 1.00 | 0.87 | 0.30 | 0.46 | 0.70 | 0.54 | C   | 0.64 | 0.60 | 0.36 | 0.43 | 0.28 | 0.43 |
| Meso       | 1.00 | 0.78 | 0.33 | 0.44 | 0.67 | 0.49 | 1.01 | 0.76 | 0.61 | 0.60 | 0.41 | 0.45 |
| T-value    | 0.06 | 0.62 | 0.56 | 3.46 | 2.62 | 1.69 | 2.62 | 1.69 | 2.62 | 1.69 | 2.62 | 1.69 |

a.a., Amino acid type; mean, the average frequency (%); SD, standard deviation; THERMO, thermophilic proteins; MESO, mesophilic proteins; and T-value, the t-test statistic value. Significant t-values (|t| > 1.96) are represented as bold, and the data discussed in the text are represented as bold and underlined.
Table 3
Preference of substitution on the protein surface

|    | G    | A    | V    | L    | I    | M    | P    | F    | Y    | W    | K    | R    | H    | D    | E    | S    | T    | N    | Q    | C    | Gap |
|----|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| G  | –    | -0.89| 0.22 | -0.33| -0.11| -0.44| -0.56| 0.11 | 0.00 | 0.33 | -0.11| 0.22 | 0.11 | 1.45 | 1.00 | -0.56| 0.00 | -0.78| -2.56| 0.22 | 2.00 |
| A  | 0.89 | –    | -0.22| 1.56 | 0.22 | -0.22| **2.22**| -0.33| 1.11 | 0.44 | **3.34**| **3.67**| -0.11| 0.56 | **4.12**| -0.44| -0.67| 1.00 | -0.44 | -0.33| 1.45 |
| V  | -0.22| 0.22 | –    | 0.11 | -0.56| -0.56| 0.44 | -0.11| 1.00 | 0.11 | 0.89 | 0.78 | -0.67 | -0.56 | 0.22 | -0.44 | -1.89 | -0.11 | -0.89 | -0.11 | 1.00 |
| L  | 0.33 | -1.56| -0.11| -   | -0.22| -1.56| 0.56 | 0.33 | 1.11 | 0.11 | -0.56| 0.56 | -0.22 | 0.00 | -0.56 | 0.22 | -0.22 | 0.11 | -0.56 | -0.44 | 0.00 |
| I  | 0.11 | -0.22| 0.56 | 0.22 | –   | -0.44| 0.67 | -1.22| 0.33 | 0.22 | -0.11| 0.00 | 0.00 | 0.00 | -0.11| 0.67 | -1.22| 0.89 | -0.22 | -0.22 | -0.11 | 2.45 |
| M  | 0.44 | 0.22 | 0.56 | 1.56 | 0.44 | –   | -0.11| -0.11| 0.00 | 0.11 | 0.33 | 0.67 | 0.22 | 0.00 | 1.22 | -0.11| -0.44| 0.00 | 0.22 | 0.00 | 0.56 |
| P  | 0.56 | -2.22| -0.44| -0.56| -0.67| 0.11 | –   | -0.44| -0.11| 0.00 | 0.11 | 0.11 | 0.56 | 1.00 | -0.56 | -1.11| -0.67 | 0.11 | -0.22 | -0.44 |
| F  | -0.11| 0.33 | 0.11 | -0.33| 1.22 | 0.11 | 0.44 | –   | 0.56 | 0.11 | 0.56 | 0.56 | -0.33 | -0.11| -0.11 | -0.11| -1.00| -0.67 | -0.44 | -0.11 | -0.33 |
| Y  | 0.00 | -1.11| -1.00| -1.11| -0.33| 0.00 | 0.11 | -0.56 | –   | -0.11| -0.78| -0.56 | -0.89 | 0.22 | -0.89 | 0.22 | -0.56 | 0.33 | -0.22 | -0.33 | 1.00 |
| W  | -0.33| -0.44| -0.11| -0.11| -0.22| -0.11| 0.00 | -0.11| 0.11 | –   | -0.22| 0.11 | 0.33 | 0.00 | -0.11| 0.00 | 0.00 | 0.00 | 0.11 | 0.00 | -0.11 |
| K  | 0.11 | -3.34| -0.89| 0.56 | 0.11 | -0.33| 0.11 | -0.56| 0.78 | -0.22| –   | 2.45 | -0.56 | 0.22 | 1.22 | -1.89| 0.67 | 1.00 | **2.22**| 0.11 | 0.22 |
| R  | -0.22| -3.67| -0.78| 0.56 | 0.00 | -0.67 | -0.11| -0.56| 0.56 | -0.11| -2.45 | –   | 1.56 | 1.45 | -0.11| 1.45 | -1.22| 0.89 | **2.56**| -0.22 | 1.67 |
| H  | -0.11| 0.11 | 0.67 | 0.22 | 0.00 | -0.22| -0.11| 0.33 | 0.89 | -0.33| 0.56 | 1.56 | –   | 0.44 | 0.78 | -0.11| -0.11| -0.11| -1.33 | -0.22 | 0.89 |
| D  | 1.45 | -0.56| 0.56 | 0.00 | 0.11 | 0.00 | -0.56 | 0.11 | -0.22| 0.00 | -0.22| -1.45 | -0.44 | –   | 3.56 | 0.56 | 0.56 | 0.00 | 1.00 | 0.00 | 2.22 |
| E  | -1.00| **4.12**| -0.22| 0.56 | -0.67| -1.22| -1.00| 0.11 | 0.89 | 0.11 | -1.22| 0.11 | -0.78 | -3.56 | –   | -1.89| -1.11| 0.00 | **3.34**| -0.33 | -1.22 |
| S  | 0.56 | 0.44 | 0.44 | -0.22| 1.22 | 0.11 | 0.56 | 0.11 | 0.22 | 0.00 | 1.89 | 1.45 | 0.11 | -0.56 | 1.89 | –   | 0.22 | 0.44 | -1.11 | -0.22 | 2.22 |
| T  | 0.00 | 0.67 | 1.89 | 0.22 | -0.89| 0.44 | 1.11 | 1.00 | 0.56 | 0.00 | -0.67 | 1.22 | 0.11 | 0.56 | 1.11 | -0.22 | –   | -1.33 | -0.44 | 0.11 | 1.00 |
| N  | 0.78 | -1.00| 0.11 | 0.11 | 0.22 | 0.00 | 0.67 | 0.67 | 0.33 | 0.00 | 1.00 | 0.89 | 0.11 | 0.00 | 0.00 | 0.00 | 0.44 | 1.33 | –   | 0.22 | 0.00 | 1.33 |
| Q  | 2.56 | 0.44 | 0.89 | 0.56 | 0.22 | 0.11 | 0.44 | 0.22 | -0.11| 0.44 | 0.22 | **2.22**| **2.56**| 1.33 | 1.00 | **3.34**| 1.11 | 0.44 | -0.22 | –   | -0.11 | 0.78 |
| C  | -0.22| 0.33 | 0.11 | 0.44 | 0.11 | 0.00 | 0.22 | 0.11 | 0.33 | 0.00 | -0.11| 0.22 | 0.22 | 0.00 | 0.33 | 0.22 | -0.11| 0.00 | 0.11 | –   | 0.11 |
| Gap| -2.00| -1.45| -1.00| 0.00 | -2.45| -0.56| 0.44 | 0.33 | -1.00| 0.11 | -0.22| -1.67 | -0.89 | -2.22| 1.22 | -2.22| 1.00 | -1.33 | -0.78 | -0.11 | –   |

The amino acids in row and column are for mesophilic and thermophilic proteins, respectively. Positive and negative signs indicate that the substitutions are favorable and unfavorable for thermophilic proteins, respectively. Significant difference data are represented as bold, and the data discussed in the text are represented as bold and underlined.
The amino acids in row and column are for mesophilic and thermophilic proteins, respectively. Positive and negative signs indicate that the substitutions are favorable and unfavorable for thermophilic proteins, respectively. Significant difference data are represented as bold, and the data discussed in the text are represented as bold and underlined.
t-values are low, show relatively high values for their replacement (the preferences are 3.67 and 2.56, respectively).

The t-value of solvent-exposed Glu clearly suggests that this amino acid type is preferentially located at the surface of thermophilic proteins. This seems to be consistent with the fact that the formation of an ion-pair or ion-pair network at the surface enhances protein thermostability as reported previously [10,14,17,24,25]. Similarly to the Arg case, Glu residues at the surface in thermophilic proteins appear by substitution of a variety of amino acids (14 amino acids), as shown in Table 3. Replacements of Ala or Gln by Glu show relatively high values (with preferences 4.12 and 3.34, respectively).

Tyr is involved in cation-pi interactions or forms aromatic clusters in protein structures, as reported previously [26,27,28]. Such properties are key factors for protein thermostability as shown in earlier works [28,29]. The present result quantitatively suggests that, together with Arg and Glu, Tyr has a tendency to exist on the surface of thermophilic proteins. The analysis of substitution preference shows that various amino acids (15 amino acids) can change to Tyr. No specificity was detected in the mutation so that this preference is slightly different from that of Arg and Glu.

3.3. Protein surface amino acids that do not favor thermostability

Analysis of the protein surface also suggests amino acids unfavorable for thermostability. As shown in Table 2, it is found that the contents of Ala, Ser, Met, and Gln at the surface are extremely low (the \( t \)-values are 3.32, 2.49, 2.12, and 3.05, respectively). The reasons for this have been discussed in other comparative works [16,30]. The present result of Ala having a lower \( t \)-value is consistent with a previous comparative analysis [17]. Moreover, our analysis indicates that the substitution of Ala by Lys, Arg, and Glu (the values of preferences are 3.34, 3.67, and 4.12, respectively) is preferred. Of course, it is thought that the lower content of Ala at the surface of thermophilic proteins is caused by its hydrophobicity, but a more important reason may be its effect on the neighboring residues, i.e. disruption of formation of ion-pairs or an ion-pair network at the protein surface. Lack of such ion-pair interactions may be a disadvantage for protein thermostability.

Ser is a dissociative amino acid with a short side chain containing one hydroxyl group. It has been proposed that Ser at the protein surface does not favor thermostability [17,25]. The present result for solvent-exposed Ser (the \( t \)-value is \(-2.12\)) quantitatively supports previous reports [17,25]. Further studies are required to address the physical mechanism of its contributes to protein instability although the reason for this observation may be related to the interaction between Ser and a water molecule. As shown in Table 3, substitutions of Ser by various amino acids (14 types with negative values) occur on the protein surface for thermostability. In particular, the substitution preferences towards Lys, Arg, and Glu are higher (1.89, 1.45, and 1.89, respectively). The results on solvent-exposed Ser are also supported by the \( t \)-value as shown in Table 2.

For thermostability, Met is believed disadvantageous when present at the protein surface. This is because the sulfur atom in its side chain is susceptible to oxidation, which would destabilize the surface structure of proteins [12]. In the present study, it is found that the content of Met at the surface of thermophilic proteins is low (the \( t \)-value is \(-2.49\)) as shown in Table 2. This is the first quantitative observation on the location preference of Met using comparative analysis [17]. Furthermore, the present study also detects that substitutions of the solvent-exposed Met by various amino acids (10 amino acids) occurred and are summarized in Table 3. This result also suggests the disadvantageousness of Met at the protein surface for thermostability.

The presence of Gln at the protein surface is thought to be disadvantageous for thermostability because Gln is associated with deamidation, which would destabilize the protein surface [11,31]. The present analysis (summarized in Table 2) indicates the lower content of Gln at the surface of thermophilic proteins (the \( t \)-value is \(-3.05\)). Thus, this is the first quantitative observation to suggest this tendency for Gln [17]. Substitution of the solvent-exposed Gln by various amino acids (15 amino acids) was detected as shown in Table 3. The substitution preferences towards Lys, Arg, and Glu are high (the preferences are 2.22, 2.56, and 3.34, respectively). Preference of these substitutions is also supported by their \( t \)-values.

3.4. Protein core amino acids that favor thermostability

We observed that the content of Ala in the core of thermophilic proteins is high (the \( t \)-value is 2.81) as shown in Table 2. Moreover, as shown in Table 4, the present study indicates that substitutions of 16 amino acid types by Ala frequently occur in the core of thermophilic proteins. In particular, the value from Ser to Ala is large (the preference is 4.98). It is thought that Ala is concerned with compact packing, which would stabilize the core of proteins [17,30]. Thus, the present results support quantitatively this hypothesis.

3.5. Protein core amino acids that do not favor thermostability

In contrast to Ala, the contents of Ser and Cys in the core of thermophilic proteins are low (the \( t \)-values are \(-2.41\) and \(-2.62\), respectively). These findings quantitatively suggest that Ser and Cys in the protein core are disadvantageous for thermostability, although they have been discussed in several works [14,16,17,25]. Our large dataset enabled investigation of such properties for Ser and Cys.

It is found that substitution of Ser by Ala often occurs (the value is 4.98) in the core of thermophilic proteins. Interestingly, in this study, lower contents of Ser in both states (i.e. surface and core) of thermophilic proteins are detected, although different reasons are expected for them as mentioned in the sections above.

Cys is a hydrophobic amino acid with a short side chain containing one sulfur atom, which is believed to be unfavorable for thermostability [16]. The reason for this is that the sulfur atom is relatively large and cannot contribute to the compact packing of protein structures. In this study, the \( t \)-value of Cys
in the core is $-2.62$ as shown in Table 2. This is the first quantitative observation of the lower content of Cys in the core of thermostable proteins [17]. From the analysis of substitution preference, we observed the substitution of Cys by nine amino acid types. More particularly, the preference for Ala has a relatively high value (1.59).

4. Conclusions

We have constructed a thermophilic–mesophilic protein homologue dataset in which the number of categories is double that of previous studies [10,14,16,17,25]. In this work, we found that (i) at the surface, a higher frequency of Arg, Glu, and Tyr, is observed and for Ala, Ser, Met, and Gln a lower frequency is observed and (ii) in the core, Ala is favored, but Ser and Cys are not. We have reported the first quantitative results with a comparative analysis especially for the preferences of Cys, Gln, Met, and Ser. The large dataset provided valuable information about the factors in protein thermostability. This kind of information is extremely helpful for an understanding of the protein thermostability. Accumulation of such knowledge will give useful hints for design of nanomaterials showing thermodynamic properties as desired.

Acknowledgements

This work was supported in part by the Center for Nano Materials and Technology, JAIST.

References

[1] P.J. Haney, J.H. Badger, G.L. Buldak, C.I. Reich, C.R. Woese, G.J. Olsen, Thermal adaptation analyzed by comparison of protein sequences from mesophilic and extremely thermophilic Methanococcus species, Proc. Natl Acad. Sci. USA 96 (1999) 3578–3583.

[2] C. Cambillau, J.M. Claverie, Structural and genomic correlates of hyperthermostability, J. Biol. Chem. 275 (2000) 32383–32386.

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

[4] A.H. Elcock, The stability of salt bridges at high temperatures: implications for hypertherophilic proteins, J. Mol. Biol. 284 (1998) 489–502.

[5] S. Kumar, B. Ma, C.J. Tsai, R. Nussinov, Electrostatic strengths of salt bridges in thermophilic and mesophilic glutamate dehydrogenase monomers, Proteins 38 (2000) 368–383.

[6] K.S. Yip, T.J. Stillman, K.L. Britton, P.J. Artymiuk, P.J. Baker, S.E. Sedelnikova, P.C. Engel, A. Pasquu, R. Chiaraluce, V. Consalvi, The structure of pyrococcus furiosus glutamate dehydrogenase reveals a key role for ion-pair networks in maintaining enzyme stability at extreme temperatures, Structure 3 (1995) 1147–1158.

[7] M. Hennig, B. Darimont, R. Sterner, K. Kirschner, J.N. Jansonius, 2.0 A structure of indole-3-glycerol phosphate synthase from the hyperthermophilic archaean pyrococcus furiosus at 1.9 A resolution, Biochemistry 36 (1997) 9983–9994.

[8] L. Xiao, B. Honig, Electrostatic contributions to the stability of hyperthermophilic proteins, J. Mol. Biol. 289 (1999) 1435–1444.

[9] M.W. Adams, R.M. Kelly, Finding and using hyperthermophilic enzymes, Trends Biotechnol. 16 (1998) 329–332.

[10] A. Szilagyi, P. Zavodszky, Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits: results of a comprehensive survey, Struct. Fold. Des. 8 (2000) 493–504.

[11] C. Vieille, G.J. Zeikus, Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability, Microbiol. Mol. Biol. Rev. 65 (2001) 1–43.

[12] R.J. Russell, J.M. Ferguson, D.W. Hough, M.J. Danzon, G.L. Taylor, The crystal structure of citrate synthase from the hyperthermophilic archaeon pyrococcus furiosus at 1.9 A resolution, Biotechnology 36 (1997) 9983–9994.

[13] G. Vogt, P. Argos, Protein thermal stability: hydrogen bonds or internal packing? Struct. Fold. Des. 2 (1997) 540–546.

[14] G. Vogt, S. Woell, P. Argos, Protein thermal stability, hydrogen bonds, and ion pairs, J. Mol. Biol. 269 (1997) 631–643.

[15] M.M. Gromiha, Important inter-residue contacts for enhancing the thermal stability of thermophilic proteins, Biophys. Chem. 91 (2001) 71–77.

[16] M.M. Gromiha, M. Obatake, A. Sarai, Important amino acid properties for enhanced thermostability from mesophilic to thermophilic proteins, Biophys. Chem. 82 (1999) 51–67.

[17] S.P. Stack, Y.J. Yoo, Protein thermostability: structure-based difference of amino acid residues between thermophillic and mesophilic proteins, J. Biotechnol. 111 (2004) 269–277.

[18] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The protein data bank, Nucleic Acids Res. 28 (2000) 235–242.

[19] A. Andreeva, D. Howorth, S.E. Brenner, T.J.P. Hubbard, C. Chothia, A.G. Murzin, SCOP database in 2004: refinements integrate structure and sequence family data, Nucleic Acids Res. 32 (2004) D226–D229 (Database issue).

[20] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (1997) 3389–3402.

[21] E. Myers, W. Miller, Optimal alignments in linear space, Comput. Appl. Biosci. 4 (1988) 11–17.

[22] A.R. Fersht, Conformational equilibria in three- and five-chymotrypsin, the energetics and importance of the salt bridge, J. Mol. Biol. 64 (1972) 497–509.

[23] S. Kumar, R. Nussinov, Relationship between ion pair geometries and electrostatic strengths in proteins, Biophys. J. 83 (2002) 1595–1612.

[24] L.I. Liu, W.C. Wang, Predicted unfolding order of the 13 alpha-helices in the catalytic domain of glucoamylase from aspergillus awamori var. X100 by molecular dynamics simulations, Biotechnol. Prog. 19 (2003) 1583–1590.

[25] S. Kumar, C.J. Tsai, R. Nussinov, Factors enhancing protein thermostability, Protein Eng. 13 (2000) 179–191.

[26] J.P. Gallivan, D.A. Dougherty, Cation-pi interactions in structural biology, Proc. Natl Acad. Sci. USA 96 (1999) 9459–9464.

[27] M.M. Flocco, L. Mowbray, Planar stacking interactions of arginine and aromatic side-chains in proteins, J. Mol. Biol. 235 (1994) 709–717.

[28] N. Kannan, S. Vishveshwara, Aromatic clusters: a determinant of thermal stability, Protein Eng. 13 (2000) 871–876.

[29] M.M. Gromiha, S. Thomas, C. Santhosh, Role of cation-pi interactions to the stability of thermophilic proteins, Prep. Biochem. Biotechnol. 32 (2002) 355–362.

[30] F. Tekia, E. Yeramian, D. Bernard, Amino acid composition of genomes, lifestyles of organisms, and evolutionary trends: a global picture with correspondence analysis, Gene 297 (2002) 51–60.

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