1 **Insight into the mechanism of protein thermostability based on the** 2 **residue interaction degrees**

3

4 Huihua Ge, Yunmeng Chu, Guangya Zhang*

5

6 Fujian Provincial Key Laboratory of Biochemical Technology (Huaqiao University),

7 Xiamen, 361021, PR China

8

9 Corresponding author:

10 Dr. Guangya Zhang

11 E-mail: zhgyghh@hqu.edu.cn; Tel: +86 592 6162302; Fax: +86 592 6162300

12 Authors email addresses:

13 Huihua Ge: 273085995@qq.com

14 Yunmeng Chu: 843705628@qq.com

15

16

17

18

19

20

21

22

1 ABSTRACT

2 Understanding the basis of protein thermostability raises a general question: which
3 residue with specific interaction degrees is more important to the protein
4 thermostability? A strictly selected dataset of 131 pairs of thermophilic (TPs) and
5 mesophilic proteins (MPs) was constructed. There were 6.4% and 8.4% of the total
6 residues in sequences did not interact with others in TPs and MPs. The amino acid
7 contents in sequences are closest to those with the interaction degrees of 3 according to
8 the Chi-squared distances. Only Glu, Gln and the amide residues showed significant
9 differences in sequences, which was the same as identified at low residue interaction
10 degrees. However, we observed significant Phe, Lys, Leu, Gln and the charged,
11 aliphatic, aromatic, positive charged and small residues at high interaction degree.
12 Among them, Phe was rarely reported previously although aromatic residues were well-
13 known contributor to protein thermostability. Finally, we took aspartate
14 transcarbamylases as an example to explain how a residue with various interaction
15 degrees contributed differently to their thermostability. Our results clearly
16 demonstrated the differences of amino acids in sequence between TPs and MPs could
17 only represent those involved in low interaction degrees. Much more residues with
18 significant differences existed at high interaction degrees even if they had few
19 significant amino acids in sequences. The interaction degree-based method should be
20 an alternative tool in extracting valuable eigenvalues for predicting proteins attributes
21 in bioinformatics. It could also provide a new perspective for studying the
22 thermostability of proteins and engineering novel thermostable proteins.

1 **Keywords:** Protein thermostability; Residue interaction degree; Feature extraction;
 2 Amino acid composition; Aspartate transcarbamylase

3

4

5

6

7 **List of abbreviations:**

8 TPs: thermophilic proteins,

9 MPs: mesophilic proteins,

10 OGT: optimal growth temperature,

11 ASA: absolute surface areas

12

13

14

15

16

17

18

19

20

21

22

1 **Introduction**

2 Elucidating the mechanisms of the stability of thermophilic proteins is an important
3 field in computational biology. Researcher have found many factors contributed to the
4 thermostability of proteins, though some of them are still debated [1-5]. However, it is
5 widely accepted amino acids play an important role in thermostabilization. Researchers
6 have observed significant differences of amino acids between TPs and MPs in the
7 proteome [6], the surface and interior of the proteins [7] and various secondary
8 structures [8]. Besides, they adopted many classifiers to discriminate thermophilic and
9 mesophilic proteins based on the amino acid composition [9-11]. The high overall
10 prediction accuracy proved the amino acid composition was an effective eigenvalue to
11 distinguish these two types of proteins.

12 However, based on the known facts, it is difficult to answer the following questions:
13 does the identical residue with different interaction degrees contribute equally to the
14 protein thermostability? If not, which residue with specific interaction degrees is more
15 important to the thermostability of proteins? Can we identify those amino acids
16 showing significant differences according to their interaction degrees? Despite many
17 works devoted to the search for amino acid differences between TPs and MPs, these
18 questions were not mentioned and solved. At least, the works existed currently did not
19 provide accurate fingerprints about the interaction degrees of amino acids when
20 compared their features between TPs and MPs.

21 From a theoretical perspective, the 3D structure of a protein was determined by a
22 delicate balance of different interactions between residues. Without a doubt, the amino

1 acids contribute to the stability of a protein *via* their mutual interactions. As we all know,
2 many interactions such as the hydrogen bonds, salt bridges, π -cationic interactions and
3 so on play important roles in the thermostability of proteins [12, 13]. Sometimes, an
4 amino acid may involve in more than one kinds of interactions in a protein, and
5 sometimes it may not interact with others in the same protein. For example,
6 thermophilic proteins showed higher frequencies of arginines in sequence or in exposed
7 state, they contributed positively to the thermostability [14, 15]. However, there are
8 many arginines in protein sequence, they have different interaction with other residues
9 depending on their positions and those residues involved. For instances, there are two
10 arginines which do not interact with others (e.g. Arg-12 and Arg-152 in 1ML4, 1ML4
11 is the aspartate transcarbamylase from the hyperthermophilic archaeon *Pyrococcus*
12 *abyssi* [16].), while three arginines have nine interactions with others (e.g. Arg-107,
13 Arg-168 and Arg-297 in 1ML4). Do they contribute equally to the protein
14 thermostability? If not, how do we differentiate their role to the thermostability of
15 proteins? Unfortunately, few works have addressed it although researcher have studied
16 protein thermostability for a long time. From an applicative perspective, the amino
17 acids with specific interaction degrees that show significant differences between TPs
18 and MPs are quite definite, especially those with high interaction degrees. Thus, it is
19 convenient to target these residues when engineering for thermophilic proteins.
20 Meanwhile, the outcomes will be improved and more trustable.

21 Here, we strictly selected thermophilic and mesophilic homologs and created a
22 dataset containing 131 pairs of proteins. They had high sequence identities and only

1 few residues showed significant differences in sequences. Our first purpose is to
2 identify more amino acids at some interaction degrees that contributed significantly to
3 the thermostability of proteins. The second purpose is to differentiate the role of the
4 same amino acid with various interaction degrees to the protein thermostability.

5 **Materials and methods**

6 **Dataset construction**

7 The original dataset of 252 pairs of TPs and MPs was compiled by Taylor *et al.* [17].
8 However, many identical proteins have two or more different mesophilic or
9 thermophilic homologues in this dataset. To reduce the computational bias, we kept
10 only one pair of them each. The criterion is a). The homologues have the higher
11 sequence identity; b). Their source organisms have higher optimal growth temperature
12 (for TPs) or have lower optimal growth temperature (for MPs); c). The sequence length
13 is closer; d). Discarding the sequences with less than 100 amino acids. We got 112 pairs
14 of homologues from this dataset. Then we gather another 19 pairs of homologues,
15 which was completely different. Finally, we got 131 pairs of proteins. The information
16 containing the PDB IDs, the length, the resolution of the structure, the source organisms,
17 their corresponding optimal growth temperature (OGT), and the sequence identity were
18 listed in the supplementary table S1. There were 40384 and 41609 amino acids in the
19 sequences of TPs and MPs. Meanwhile, there were 2578 and 3481 amino acids did not
20 interact with others, accounting 6.4% and 8.4% for TPs and MPs.

21 **Interaction degree**

22 The interactions calculated here include hydrogen bond, Van der Waals, disulphide

1 bond, salt bridge, π - π stacking and π -cation according to the Residue Interaction
 2 Network Generator (RING) at the RING 2.0 web server [18]. The cutoff distances for
 3 hydrogen bond, Van der Waals, disulphide bond, salt bridge, π - π stacking and π -cation
 4 were 3.5 Å, 0.5 Å, 2.5 Å, 4.0 Å, 6.5 Å and 5.0 Å, respectively [18]. The interaction
 5 degree (D) of an amino acid means the number of interactions it interacts with others.
 6 We divided the degrees into three groups, namely, low degree ($1 \leq D \leq 4$), medium
 7 degree ($5 \leq D \leq 8$) and high degree ($D \geq 9$). Their proportion was 59.1%, 32.4% and
 8 2.1% for TPs, while 59.4%, 30.3% and 1.9% for MPs, respectively.

9 Definition of the Chi-squared distances

10 We might count the frequency of each residue at a specific interaction degree and
 11 compared them with what we would expect (the frequency of each residue in the
 12 sequence) by means of Chi-squared distances. If there were no preference of the
 13 residues at a specific interaction degree, the frequencies of the residues at the specific
 14 interaction degree (observed) would be the same as those in the sequence (expected).
 15 Thus, the distances would be zero. Therefore, the Chi-squared distances can help us
 16 make decisions about whether the observed outcome differs significantly from the
 17 expected outcome.

18 The Chi-squared distances between sequence and different interaction degrees was
 19 defined as following according to Abdullah's work [19]:

$$20 \quad CSD = \sum_{k=1}^{20} \frac{(x_{d,k} - x_{s,k})^2}{(x_{d,k} + x_{s,k})}$$

21 Where k is the 20 amino acids in proteins, $x_{d,k}$ is the average amino acid

1 composition at a specific interaction degree, $x_{s,k}$ is the average amino acid
2 composition in sequences.

3 The critical value for the chi-square in this case is 30.144 (with the degree of
4 freedom of 19). If the calculated Chi-square value is equal to or greater than this critical
5 value, we can conclude that the probability of the null hypothesis being correct is 0.05
6 or less, which is a very small probability indeed! Then we reject the null hypothesis.
7 For example, our calculated value of 79.511(the distances between sequence and the
8 interaction degree of 10 for TPs, as shown in figure 2.) is greater than the critical value
9 of 30.144. We therefore reject the null hypothesis, and conclude that there is a
10 significant difference in amino acid composition between the sequence and residues at
11 the interaction degree of 10 in TPs.

12 Similarly, for the Chi-squared distances between sequence and different secondary
13 structure, the distances were defined as:

$$14 \quad CSD = \sum_{k=1}^{20} \frac{(x_{c,k} - x_{s,k})^2}{(x_{c,k} + x_{s,k})}$$

15 Where k is the 20 amino acids in proteins $x_{c,k}$ is the average amino acid
16 composition at a specific secondary structure (α -helix, β -sheet and coil), $x_{s,k}$ is the
17 average amino acid composition in sequences.

18 For the Chi-squared distances between the different relative solvent accessibility
19 and the sequences, the distances were defined as:

$$20 \quad CSD = \sum_{k=1}^{20} \frac{(x_{r,k} - x_{s,k})^2}{(x_{r,k} + x_{s,k})}$$

Where k is the 20 amino acids in proteins, $x_{r,k}$ is the average amino acid composition at a specific relative solvent accessibility state. $x_{s,k}$ is the average amino acid composition in sequences. According to relative solvent accessibility, the states of amino acids in proteins were classified as being in one of the four classes: completely buried (0-4% exposed), partly buried (4-25% exposed), partly exposed (25-50% exposed) and completely exposed (50+% exposed) [20].

Significant amino acids (or amino acid groups) detection

The differences of amino acid (or amino acid groups) composition between TPs and MPs were defined as:

$$D_i = \frac{(y_{i,T} - y_{i,M})^2}{(y_{i,T} + y_{i,M})}$$

Where, $y_{i,T}$ and $y_{i,M}$ are the amino acid (or amino acid groups) composition of TPs and MPs, i means the specific degree (or degree range) or structure (e.g. the secondary structure such as helix, sheet or coil).

If $D_i > 0.2$, y_i is identified as significant amino acid (or amino acid group). We classified the amino acid as the following groups according to their physicochemical properties [21, 22]. The amino acid groups include the charged, aliphatic, aromatic, polar, neutral, and hydrophobic, positive charged, negative charged, tiny, small, large, sulfur and the amide.

Results and Discussion

The Chi-squared distances between the sequence and interaction degrees based on amino acid composition

1 To see if the Chi-squared distances can reflect the amino acids preferences from
2 sequences, we calculated the Chi-squared distances between sequence and various
3 secondary structure or relative solvent accessibility of TPs and MPs and showed the
4 results in figure 1. We observed the Chi-squared distances of various secondary
5 structure and the partial exposed and buried were between 2.5 and 12, while the
6 distances of the exposed and buried were between 20 and 24. Among them, the exposed
7 state shows the most distinct differences between TPs and MPs. This means the amino
8 acid contents in the exposed part (protein surface) is the most biased when compared
9 with those in the sequences. The Chi-squared distances were 23.55 and 21.38 for TPs
10 and MPs. The absolute difference value was 2.17, which showed the most obvious
11 differences between TPs and MPs. Researchers have found the most of the amino acid
12 composition bias between TPs and MPs came from the protein surface (or the exposed
13 part) [4, 7, 23]. Our results confirm it by the precise data of Chi-squared distances. As
14 for the secondary structure, no significant differences observed between TPs and MPs
15 in terms of Chi-squared distances, which is also in accordance with previous result [8,
16 24]. The amino acids in the β -sheet shows the maximum deviation over the sequence,
17 while α -helix the minimum. The Chi-squared distances of β -sheet is 9.17 and 8.4 in TPs
18 and MPs, while 4.34 and 4.15 for α -helix, respectively. However, the distances
19 difference of the coil is more obvious between TPs and MPs. The distances are 6.33
20 and 5.36, with an absolute difference value of 0.98.

21 As the Chi-squared distances is the summation value of the distances for each of
22 the 20 amino acids, it can also identify the amino acid that contributed the greatest to

1 the distances. The residue contributed the greatest to the distances is the most biased
2 one when compared with its content in sequence. For instances, the primary contributor
3 to the distances between the protein surface and sequence is Leu in both TPs and MPs
4 (accounting 16.9% and 16.6%), as the content of Leu is much less in the exposed part
5 than in the sequences. The primary contributor for the distances between the buried part
6 and sequence is Lys in both TPs and MPs (accounting 19.2% and 20.1%); as the content
7 of Lys is much less in the buried part. As we all know, Leu is a hydrophobic residue,
8 which involves in the hydrophobic interactions in the buried part of a protein;
9 meanwhile, Lys is a polar charged residue, which should be abundant in the exposed
10 part of a protein [4]. Thus, they become the most biased residues in the exposed and
11 buried parts. For the secondary structure, Gly is the primary contributor to the distances
12 between the α -helix and sequence, and Val the primary contributor between the β -sheet
13 and sequence. According to Chou & Fasman's secondary structure parameters, Gly is
14 the strongest α -helix breaker, while Val the strongest β -sheet admirer [25]. Therefore,
15 much less Gly in α -helix and more Val in β -sheet was observed and they became the
16 primary contributor to the Chi-squared distances. Based on these results, we may
17 conclude the Chi-squared distance we proposed here is biologically meaningful and
18 convincing. Next, we will analyze the Chi-squared distances between the sequence and
19 various interaction degrees.

20 We computed the average amino acid composition of the sequence and various
21 interaction degrees and calculated the Chi-squared distances between them as showed
22 in figure 2. We can see both TPs and MPs have very similar trends. The distances

1 decreased before degree 3 and increased gradually after degree 4. The distances were
2 the minimal when the degree is 3 and 4. They were 2.12 and 2.49 for TPs, while 2.00
3 and 2.16 for MPs, respectively. This means the amino acid composition at degree 3 and
4 4 approximates that of the sequence. The most important contributor to the distances
5 comes from Phe (account 34.0% and 26.0% for TPs and MPs) at degree 3 and Gly
6 (account 52.2% and 49.1% for TPs and MPs) at degree 4. Besides, the distance values
7 of TPs were smaller than those of MPs when the degree is less than or equal to 6, and
8 they became larger when the degree is greater or equal to 7. This means the usage of
9 amino acids at relatively high degree of MPs biased more from their sequences. When
10 we sorted the interaction degrees into low, medium and high, the Chi-squared distances
11 of TPs and MPs were 66.49 vs 75.43 at high interaction degree, which was greater than
12 the critical values of 30.144. This means there is a significant difference in amino acid
13 composition between the sequence and residues at the high interaction degree. However,
14 the Chi-squared distances at low (3.88 vs 3.40) and medium degree (17.77 vs 17.98)
15 were less than the critical values, which means no significant differences observed. The
16 most important contributors to the distances were Phe (account 18.0% and 21.2% for
17 TPs and MPs) at high degree, while Gly (account 33.9% and 30.6%) at medium degree
18 and Phe (account 19.8% and 21.2%) at low degree, respectively. Interestingly, the
19 content of Phe in sequence is less than that of high degree, while more than low degree.
20 As we know, Phe has a benzene-like ring structure in its side chain and it is a highly
21 hydrophobic amino acid. It could involve in many kinds of interactions for maintaining
22 conformational stability in protein structure [26]. Consequently, the frequency of Phe

1 in high interaction degrees is much higher than that in sequence.

2 If a residue in the sequence did not interact with others, they should scarcely have
3 contributions to the thermostability of proteins. We identified those residues who did
4 not involve in the interaction and calculated the contents of them in TPs and MPs. We
5 found the content was 6.4% and 8.4% for TPs and MPs, respectively. The content is 2%
6 less in TPs, meaning more amino acids in TPs involve in the interactions with other
7 residues. The Chi-squared distances of TPs and MPs between sequence and the non-
8 involved residues (amino acids do not involve in the interactions) were shown in figure
9 1. The distances for TPs and MPs were 23.16 and 15.96, respectively, which is also less
10 than the critical value.

11 When compared the Chi-squared distances of the high interaction degree with
12 those of the exposed or buried part (as shown in figure 1 and 2), we found the distance
13 of the high interaction degree was much more. This means the usage of the amino acids
14 at high interaction degree is more biased than that on the exposed and buried part of
15 TPs and MPs. As mentioned above, the exposed and buried part were distinctively differ
16 between TPs and MPs, which has been proved by many earlier researchers [1, 2, 4, 5,
17 7]. Here, we found the amino acids were far more biased at the high interaction degrees,
18 we believe the differences between TPs and MPs are extremely significant at high
19 interaction degree when compared with the surface and core part of proteins. Next, we
20 will discuss the amino acid differences between the TPs and MPs according to their
21 interaction degrees in detail.

22 **The amino acids showing significant differences between TPs and MPs at various**

1 interaction degrees

2 To prove our definition about the significant amino acids was effective, we firstly listed
3 the amino acids showed significant differences between TPs and MPs in sequences and
4 various relative solvent accessibility and secondary structure in table 1. We found there
5 were significantly more Glu while fewer Gln and amide residues in the sequences of
6 TPs. This is consistent with previous results [15, 27]. As for the relative solvent
7 accessibility, we found the most obvious differences were in the exposed parts. There
8 were significant more Glu, Arg, charged, negative charged and large residues while less
9 Ala, Ser, neutral, tiny and amide residues in the TPs. However, we only detected
10 significant less Cys, Gln, and sulfur-containing and amide residues in the buried and
11 partial buried area. This indicates more significant amino acids between TPs and MPs
12 come from the external surface of proteins. These results were in good accordance with
13 many previous reports that the differences were the most notable in protein surface,
14 whereas not distinctive in the interior parts [4, 5, 7, 24, 28]. As for the secondary
15 structure, we found the most obvious differences in the β -sheet. There were significant
16 less Cys, Gln, Ser, tiny, sulfur-containing and amide residues in β -sheets. Meanwhile,
17 we also observed significant more Glu, less Gln, tiny, amide residues in α -helix and
18 only significant more Glu, and less amide residues in the coil. Most of our results were
19 consistent with previous researches [8, 24, 28], although they were not as detailed as
20 ours. From above-mentioned results, it is reasonable to believe the reliability of our
21 definition on the significant amino acids (or amino acid groups) between TPs and MPs.
22 Therefore, we can adopt it to identify the significant amino acids (or amino acid groups)

1 between TPs and MPs according to their interaction degrees.

2 We also listed the significant amino acids and amino acid groups at various
3 interaction degrees in table 1. As can be seen that we detected more significant amino
4 acid and amino acid groups when compared with the sequence. Among them, the
5 significant residues and amino acid groups at degree 3 and 4 are nearly the same as
6 sequence. As mentioned above, the Chi-squared distances between sequence and degree
7 3 and 4 are the minimum (Fig. 2). Taking these results together, we may conclude the
8 amino acids differences in sequence between TPs and MPs mostly represented those
9 who have three or four interactions with other residues.

10 For the significant amino acids, six of the 20 residues were assigned as significant
11 (account 30% of the total) when the interaction degrees were eight and ten. Meanwhile,
12 six of the 13 amino acid groups (with a proportion of 46.2%) also showed significant
13 differences between TPs and MPs at the degree of 10 or above. As we strictly selected
14 the homologous TPs and MPs, and only observed notable more Glu and less Gln in the
15 sequences of TPs, these two residues also appeared in most of the interaction degrees
16 as shown in table 1. In contrast, those significant amino acids detected at some
17 interaction degrees could not find in the sequences. For example, less Ser and Thr in
18 TPs at the interaction degree of 1; less Thr while more Arg in TPs at 2; and more Lys,
19 Arg and less Asp in TPs at 9, respectively. Some previous works also identified these
20 residues by comparing the sequences of TPs and MPs [8, 15, 24, 28]. These results
21 indicated the method based on the interaction degree was more sensitive and could
22 provide precise fingerprints about the significant amino acids (or groups). What needs

1 to point out is that some amino acids play opposite roles at different interaction degrees.
 2 For instances, more Tyr in TPs at the interaction degree of 6 while less Tyr at 8; more
 3 Ala in TPs at 10 while less at 8; and more Thr in TPs while less at 1 and 2. These
 4 residues should be the potential mutation points when engineering the proteins for more
 5 stable at high temperature.

6 We also calculated the differences of amino acids between TPs and MPs according
 7 to the various ranges of interaction degrees and showed them in figure 3. We identified
 8 seven significant amino acids between TPs and MPs at various ranges of interaction
 9 degrees. Among them, two residues came from the low interaction degree, one from the
 10 medium degree and four from the high degree. At the low interaction degree, obvious
 11 more Glu and less Gln existed in TPs, which is the same as detected in the sequences.
 12 At the medium interaction degree, Cys was significantly lower in TPs, which has also
 13 reported before [15]. Interestingly, most of the significant amino acids were
 14 concentrated at the high interaction degree. We detected significantly more Phe, Lys,
 15 Leu and less Gln in TPs. More Lys, Leu and less Gln in TPs were commonly known [1,
 16 8, 15] although most of them were not detected in the sequence. However, more Phe in
 17 TPs at high interaction degree borne out here is different. Although aromatic residues
 18 contribute positively to the conformational stability of TPs, only Tyr was frequently
 19 reported to have higher content in TPs [8, 15, 23, 24, 29]. As a contrast, rare reports
 20 about Phe contributing positively to the thermostability of proteins before. As we know,
 21 Phe is an aromatic residue, it can involve in the π - π stacking and π -cation in the aromatic
 22 clusters [30]. Moreover, the interaction between Phe and Cys/Met side chain at

positions ($i, i \pm 4$) can stabilize helices through interaction between aromatic electrons and sulfur atoms [31]. As thermophilic proteins usually have shorter sequence length [32] while more interactions [33], it is reasonable to expect the amino acids (such as Phe) involving in many interactions would be favorable in TPs. Some experimental results also confirmed the important role of Phe in the thermostability of proteins. For example, the T4 lysozyme mutant Ser 117 $\rightarrow$ Phe was isolated fortuitously and found to be more thermostable than wild-type by 1.1–1.4 kcal/mol [34]. While, the perturbation of the central residue (Phe 251) reduced the stability of the native structure of β -glucosidase [35].

We also calculated the differences of amino acids groups between TPs and MPs and showed them in figure 4. From it, we detected eight significant groups altogether, two groups came from the low interaction degree, one from the medium and five from the high. About 62.5% of the significant amino acid groups gathered at high interaction degrees. They were charged, positive charged, aliphatic, aromatic and small residues. It is well known these amino acid groups correlated positively to the protein thermostability [1, 2, 4, 8, 15, 23, 24, 29]. However, we could not detect them in sequences with the same dataset. As mentioned earlier, the Chi-squared distances between sequence and the low interaction degree were the minimum (Fig. 2). Combining the two results, we can deduce the differences of amino acids in sequences between TPs and MPs could only represent those involved in low interaction degree and could not represent those high. Unfortunately, the amino acids involved in high interaction degrees are usually crucial to proteins. Thus, focus on the differences of the

1 amino acids with high interaction degrees between TPs and MPs is of more importance
2 and biological significance.

3 Finally, as mentioned above, there were about 6.4% and 8.4% of the total amino
4 acid in TPs and MPs did not involve in the interaction with other residues in our datasets.
5 These non-involved residues may become the noise data when researchers tried to
6 detect differences in sequences between TPs and MPs. Thus, this sometimes will lead
7 to the conflict of their results. For example, some researcher observed more Gly in TPs
8 while others found less Gly in TPs [27, 29]. Maybe because Gly is the most abundant
9 residue among those who did not interact with other residues. As we found Gly
10 accounted for 17.2% and 15.0% of the total residues in TPs and MPs that did not involve
11 in the interaction with other residues. Besides, the interaction degree based method
12 could remove the noisy amino acids in the sequence; it could be an important auxiliary
13 tool for extracting features for predicting protein attributes [36], such as subcellular
14 localization prediction, protein structural class, and membrane protein prediction and
15 so on.

16 In short, detecting the significant amino acids between TPs and MPs based on their
17 corresponding interaction degrees not only removed the noisy data in the sequences but
18 also found more residues crucial for the thermostability of proteins. It could also
19 distinguish the contribution of specific amino acid with various interaction degrees to
20 the thermal stability of the protein. We will discuss this in the next paragraph.

21 **Differences of the interaction networks between the thermophilic and**
22 **mesophilic aspartate transcarbamylase**

1 To explain how a residue with various interaction degrees contributes differently to the
2 thermostability of proteins according to their interaction network, we took thermophilic
3 and mesophilic aspartate transcarbamylase as an example. The thermophilic one (PDB
4 id: 1ML4) came from *Pyrococcus abyssi*, with an optimal growth temperature (OGT)
5 of 97°C [16]. The mesophilic one (PDB id: 1Q95) came from *Escherichia coli*, with an
6 OGT of 37°C [37]. The reason we chose them was that they shared 52.3% sequence
7 identity and were close in sequence length (308 vs 310). The interactions (edges) within
8 the molecule were 602 and 596 for 1ML4 and 1Q95 as shown in figure 5 A and B
9 (drawn by Cytoscape [38]). The topology of the network was a little bit different. 1ML4
10 (TPs) has more compacted interaction in some area, while the interaction within 1Q95
11 was sparser and no obvious dense area detected (Fig. 5 A and B). Moreover, they also
12 shared analogous structure according to the global alignment based on the interaction
13 network as shown in figure 5 C and D. On the other hand, there were about 293 and
14 290 residues involved in the interaction, accounting 95.1% and 93.5% of the total
15 residues (as shown in table 2). About 4.9% and 6.9% of the total residues did not involve
16 in the interaction (the ratio is 6.4% and 8.4% for TPs and MPs in the whole dataset).
17 This suggested more residues in MPs did not interact with others. These non-involved
18 residues could be the potential mutations to increase the interactions, thus improving
19 its thermostability. The residue content at low interaction degree is near, while the 1Q95
20 (mesophilic) have more residues with medium degree. Nevertheless, the differences of
21 residue content at the high interaction degree are significant with $D_i = 0.78$. Eight
22 residues with high interaction degree exist in 1ML4. They include Met-66, Tyr-226,

1 Arg-297, Arg-107, Tyr-144, Arg-168, Met-265 and Met-298. However, only three in
2 1Q95, including Arg-105, Tyr-226 and Gln-288.

3 We adopted the MCODE [39] to find the highly interconnected regions that the
4 residues with high interaction degrees involved and showed them in figure 6. We found
5 five highly interconnected regions in 1ML4 while only two in 1Q95. This is due to
6 more residues with high interaction degree in 1ML4. Besides, more residues with low
7 or medium interaction degrees participated in these regions, leading more complex and
8 concentrated interaction networks in 1ML4. Figure 6A showed the sub-network with
9 the highest score (4.542) in 1ML4, there were 109 interactions (edges) and 49 residues
10 (nodes) in the network, including the Tyr-226, Arg-297, Arg-107 and Arg-168 with high
11 interaction degrees. Figure 6B showed the network with the second high score (4.0),
12 there were 22 interactions and 12 residues in it, including the Met-66, Arg-297 and Met-
13 298 with high interaction degrees. In the five sub-network of 1ML4, Arg-297 appeared
14 3 times, which was the most frequent residue. Undoubtedly, this Arg at position 297,
15 which has the interaction degree of 10, plays a critical role in the thermostability of
16 1ML4. According to the annotation in PDB, Arg-297 located at the opposite of the
17 binding pocket bottom, while Arg-107 and Arg-168 were part of the binding pocket. As
18 mentioned in the introduction part, there were two argines (Arg-12 and Arg-152) not
19 interacting with other residues in 1ML4, accounting 9.5% of the total argines. Their
20 contributions to the thermostability of 1ML4 are undoubtedly different from Arg-107,
21 Arg-168 and Arg-297, which have the interaction degrees of 9, 9 and 10, respectively.
22 In contrast, figure 6F showed the sub-network with the highest score (3.0) in 1Q95.

1 There were only 21 interactions and 15 residues in the network, including the Arg-105
 2 with high interaction degrees. Unlike 1ML4, all the argines in 1Q95 take part in the
 3 interaction with other residues, of which 10 argines come from low interaction degree,
 4 4 from medium and 1 from high. Obviously, these argines contribute differently to the
 5 stability of 1Q95. Based on these facts, we firmly believe, the result will be more
 6 accurate and reliable if we remove the non-involved argines and compared the rest
 7 argines according to their interaction degrees. So do as other residues. Thus, we can
 8 detect more amino acids that are important when pursuing the influence of amino acids
 9 on the thermostability of proteins.

10 Moreover, figure S1 compares the interaction degree and evolutionary
 11 conservation score [40] of each residue in 1ML4 and 1Q95. There is a negative
 12 correlation between interaction degree and conservation score ($r=0.42$ and 0.43 ,
 13 respectively), meaning the amino acid with higher interaction degree remains more
 14 conserved. Generally, the conserved residues are less flexible and they are usually in
 15 the buried or surface region, whereas the disorder residues are the least conserved [41].
 16 This has been proved by the results that both the hydrophobic core packing and the
 17 external residues were key factors for the stability of TPs [1, 2, 4, 7]. Meanwhile, most
 18 of the residues with high interaction degree are highly conserved. Their relative values
 19 of absolute surface areas (ASA) [42] were also the least, although some residues with
 20 low or medium degrees also have the least ASA (as shown in figure S2). About 81.8%
 21 of the non-involved amino acids (with the interaction degree of 0) have the ASA above
 22 0.5, indicating these residues are mainly located on the surface of the aspartate

1 transcarbamylase. They should be the first choice of mutation to improve the
2 thermostability of this enzyme.

3 **Conclusion**

4 In summary, we strictly selected 131 pairs of TPs and MPs, and detected the amino
5 acids that showed significant differences based on their interaction degrees. We found
6 more amino acids in the sequences of MPs did not interact with others; they might be
7 the noisy data when identify the significant amino acids from sequences. Thus leading
8 to obtain different sets of scarce and abundant amino acids for different research. For
9 those residues involved in the interactions, the more interaction degrees of the residue,
10 the more conserved of the amino acid, and the more important of it in maintaining the
11 thermostability of proteins. We could observe much more significant amino acids (or
12 amino acid groups) with high interaction degrees than in the sequences, indicating the
13 interaction degree based method was more sensitive. We found significantly more Phe
14 in TPs at high interaction degrees in the condition of no significant differences of Phe
15 content existing in the sequences with the same dataset. As the interaction degree-based
16 method can remove the noise of the sequence, it may act as an alternative tool to extract
17 effective features to predict protein attributes in bioinformatics. Meanwhile, the
18 residues with exact interaction degrees are more specific, which can provide more
19 accurate and guiding information in engineering more thermostable proteins.

20

21

22

1

2 **Declarations:**

3 **Ethics approval and consent to participate**

4 Not applicable

5 **Consent for publication**

6 Not applicable

7 **Availability of data and material**

8 All data generated or analyzed during this study are included in this published article

9 [and its Additional file].

10 **Competing interests**

11 The authors declare that they have no competing interests.

12 **Funding**

13 This work was supported by the national natural science foundation of China (No.

14 21376103).

15 **Authors' contributions**

16 Conceived and designed the experiments: HG YC GZ. Performed the experiments: HG YC. Analyzed

17 the data: HG YC GZ. Wrote and revised the paper: HG and GZ.

18 **Acknowledgements**

19 Not applicable

20

21

22

References

1. Gromiha MM, Pathak MC, Saraboji K, Ortlund EA, Gaucher EA. Hydrophobic environment is a key factor for the stability of thermophilic proteins. *Proteins*. 2013; 81(4): 715-721.
2. Tych KM, Batchelor M, Hoffmann T, Wilson MC, Hughes ML, Paci E, Brockwell DJ, Dougan L. Differential effects of hydrophobic core packing residues for thermodynamic and mechanical stability of a hyperthermophilic protein. *Langmuir*. 2016; 32(29): 7392-7402.
3. Sammond DW, Kastelowitz N, Himmel ME, Yin H, Crowley MF, Bomble YJ. Comparing residue clusters from thermophilic and mesophilic enzymes reveals adaptive mechanisms. *Plos One*. 2016; 11(1): e0145848.
4. Glyakina AV, Garbuzynskiy SO, Lobanov MY, Galzitskaya OV. Different packing of external residues can explain differences in the thermostability of proteins from thermophilic and mesophilic organisms. *Bioinformatics*. 2007; 23(17): 2231-2238.
5. Sælensminde G, Halskau Jr Ø, Jonassen I. Amino acid contacts in proteins adapted to different temperatures: hydrophobic interactions and surface charges play a key role. *Extremophiles*. 2009; 13(1): 11-20.
6. Zeldovich KB, Berezovsky IN, Shakhnovich EI. Protein and DNA sequence determinants of thermophilic adaptation. *Plos Comp Bio*. 2007; 3(1): e5.
7. Fukuchi S, Nishikawa K. Protein surface amino acid compositions distinctively differ between thermophilic and mesophilic bacteria. *J Mol Biol*. 2001; 309(4): 835-843.

- 1 8. Pack SP, Kang TJ, Yoo YJ. Protein thermostabilizing factors: high relative
2 occurrence of amino acids, residual properties, and secondary structure type in
3 different residual state. *Appl Biochem Biotechnol*. 2013; 171(5): 1212-1226.
- 4 9. Gromiha MM, Suresh MX. Discrimination of mesophilic and thermophilic proteins
5 using machine learning algorithms. *Proteins*. 2008; 70(4): 1274-1279.
- 6 10. Ebrahimi M, Lakizadeh A, Agha-Golzadeh P, Ebrahimie E, Ebrahimi M. Prediction
7 of thermostability from amino acid attributes by combination of clustering with
8 attribute weighting: a new vista in engineering enzymes. *Plos One*. 2011; 6(8):
9 e23146.
- 10 11. Zhang G, Fang B. Support vector machine for discrimination of thermophilic and
11 mesophilic proteins based on amino acid composition. *Pro Pep Lett*. 2006; 13(10):
12 965-970.
- 13 12. Robinson-Rechavi M, Alibés A, Godzik A. Contribution of electrostatic
14 interactions, compactness and quaternary structure to protein thermostability:
15 lessons from structural genomics of *Thermotoga maritima*. *J Mol Biol*. 2006;
16 356(2): 547-557.
- 17 13. Ibrahim BS, Pattabhi V. Role of weak interactions in thermal stability of proteins.
18 *Biochem Biophys Res Commun*. 2004; 325: 1082–1089.
- 19 14. Cambillau C, Claverie J M. Structural and genomic correlates of
20 hyperthermostability. *J Biol Chem*. 2000; 275(42): 32383-32386.
- 21 15. Zhou XX, Wang YB, Pan YJ, Li WF. Differences in amino acids composition and
22 coupling patterns between mesophilic and thermophilic proteins. *Amino Acids*.

- 1 2008; 34(1): 25-33.
- 2 16. Van Boxtael S, Cunin R, Khan S, Maes D. Aspartate transcarbamylase from the
3 hyperthermophilic archaeon *Pyrococcus abyssi*: thermostability and 1.8 Å
4 resolution crystal structure of the catalytic subunit complexed with the bisubstrate
5 analogue N-phosphonacetyl-L-aspartate. *J Mol Biol.* 2003; 326(1): 203-216.
- 6 17. Taylor TJ, Vaisman II. Discrimination of thermophilic and mesophilic proteins.
7 BMC Struct Biol; 2010; 10(1): 1.
- 8 18. Piovesan D, Minervini G, Tosatto SCE. The RING 2.0 web server for high quality
9 residue interaction networks. *Nucleic Acids Res.* 2016; gkw315.
- 10 19. Abdullah MFA, Sayeed MS, Muthu KS, Bashier HK, Azman A, Ibrahim SZ. Face
11 recognition with symmetric local graph structure (slgs). *Expert Syst Appl.* 2014;
12 41(14): 6131-6137.
- 13 20. Mooney C, Pollastri G. Beyond the twilight zone: automated prediction of
14 structural properties of proteins by recursive neural networks and remote homology
15 information. *Proteins.* 2009; 77(1): 181-190.
- 16 21. Nath A, Subbiah K. Insights into the molecular basis of piezophilic adaptation:
17 Extraction of piezophilic signatures. *J Theo Biol.* 2016; 390: 117-126.
- 18 22. Jahandideh S, Asadabadi EB, Abdolmaleki P, Jahandideh M, Hoseini S. Protein
19 psychrophilicity: role of residual structural properties in adaptation of proteins to
20 low temperatures. *J Theo Biol.* 2007; 248(4): 721-726.
- 21 23. Pack SP, Yoo YJ. Protein thermostability: structure-based difference of amino acid
22 between thermophilic and mesophilic proteins. *J Biotechnol.* 2004; 111(3): 269-

- 1 277.
- 2 24. Szilágyi A, Závodszy P. Structural differences between mesophilic, moderately
3 thermophilic and extremely thermophilic protein subunits: results of a
4 comprehensive survey. *Structure*. 2000; 8(5): 493-504.
- 5 25. Chou PY, Fasman GD. Conformational parameters for amino acids in helical, β -
6 sheet, and random coil regions calculated from proteins. *Biochemistry*. 1974; 13(2):
7 211-222.
- 8 26. Dougherty D A. The cation- π interaction. *Accounts Chem Res*. 2012; 46(4): 885-
9 893.
- 10 27. Haney PJ, Badger JH, Buldak GL, Reich CI, Woese CR, Olsen GJ. Thermal
11 adaptation analyzed by comparison of protein sequences from mesophilic and
12 extremely thermophilic *Methanococcus* species. *Proc Natl Acad Sci USA*. 1999;
13 96(7): 3578-3583.
- 14 28. Kumar S, Tsai CJ, Nussinov R. Factors enhancing protein thermostability. *Protein*
15 *Eng*. 2000; 13(3): 179-191.
- 16 29. Farias ST, Bonato MC. Preferred amino acids and thermostability. *Genet Mol Res*.
17 2003; 2(4): 383-393.
- 18 30. Kannan N, Vishveshwara S. Aromatic clusters: a determinant of thermal stability
19 of thermophilic proteins. *Protein Eng*. 2000; 13(11): 753-761.
- 20 31. Chakravarty S, Varadarajan R. Elucidation of factors responsible for enhanced
21 thermal stability of proteins: a structural genomics based study. *Biochemistry*. 2002;
22 41(25): 8152-8161.

- 1 32. Brocchieri L, Karlin S. Protein length in eukaryotic and prokaryotic proteomes.
2 *Nucleic Acids Res.* 2005; 33(10): 3390-3400.
- 3 33. Gromiha MM. Important inter-residue contacts for enhancing the thermal stability
4 of thermophilic proteins. *Biophys Chem.* 2001; 91(1): 71-77.
- 5 34. Anderson DE, Hurley JH, Nicholson H, Baase WA, Matthews BW. Hydrophobic
6 core repacking and aromatic–aromatic interaction in the thermostable mutant of T4
7 lysozyme Ser 117→ Phe. *Protein Sci.* 1993; 2(8): 1285-1290.
- 8 35. Souza VP, Ikegami CM, Arantes GM, Marana SR. Protein thermal denaturation is
9 modulated by central residues in the protein structure network. *FEBS J.* 2016; 283:
10 1124-1138.
- 11 36. Chou K C. Some remarks on protein attribute prediction and pseudo amino acid
12 composition. *J theo biol*, 2011, 273(1): 236-247.
- 13 37. Huang J, Lipscomb WN. Aspartate transcarbamylase (ATCase) of *Escherichia coli*:
14 a new crystalline R-state bound to PALA, or to product analogues citrate and
15 phosphate. *Biochemistry.* 2004; 43(21): 6415-6421.
- 16 38. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N,
17 Schwikowski B, Ideker T. Cytoscape: a software environment for integrated
18 models of biomolecular interaction networks. *Genome Res.* 2003; 13(11): 2498-
19 2504.
- 20 39. Bader GD, Hogue CWV. An automated method for finding molecular complexes
21 in large protein interaction networks. *BMC bioinformatics.* 2003; 4(1): 1.
- 22 40. Ashkenazy H, Erez E, Martz E, Pupko T, Ben-Tal N. ConSurf 2010: calculating

1 evolutionary conservation in sequence and structure of proteins and nucleic acids.

2 *Nucleic Acids Res.* 2010; gkq399.

3 41. Marsh JA, Teichmann SA. Parallel dynamics and evolution: protein
4 conformational fluctuations and assembly reflect evolutionary changes in sequence
5 and structure. *BioEssays.* 2014; 36(2): 209-218.

6 42. Shander A, Gromiha M, Fawareh H Sarai A. ASA view: solvent accessibility
7 graphics for proteins. *Bioinformatics.* 2004; 51: 51.

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

1

2 **Figure Captions**

3 **Fig. 1** The Chi-squared distances between sequence and secondary structure, relative
4 solvent accessibility and non-involved amino acids of TPs and MPs.

5 **Fig. 2** The Chi-squared distances between sequence and different residue interaction
6 degrees.

7 **Fig. 3** The differences of amino acids between thermophilic and mesophilic at various
8 residue interaction degrees.

9 The upwards arrow means more in TPs, and the downwards arrow means more in MPs.

10 **Fig. 4** The differences of amino acid groups between TPs and MPs at various residue
11 interaction degrees.

12 The upwards arrow means more in TPs, and the downwards arrow means more in MPs.

13 **Fig. 5** The residual interaction network of 1ML4 and 1Q95.

14 A, B: Circular diagram representation of 1ML4 and 1Q95 interaction network; residues
15 are presented as blue dots evenly distributed around the circle in a clockwise order
16 starting from the black bar, which indicates the first residue involved in interaction with
17 degree 1. Interaction between residues are represented by black lines.

18 C, D: the global alignment of 1ML4 and 1Q95 based on residue interaction network.

19 The top 10 largest common collected subgraphs (CCSs) are colored differently.

20 **Fig. 6** The highly interconnected regions the residues with high degrees (≥ 9) involved
21 in 1ML4 and 1Q95.

22 A-E: regions in 1ML4. The high degree residues involved include Met-66, Tyr-226,

1 Arg-297, Arg-107, Tyr-144, Arg-168, Met-265 and Met-298.
 2 F-G: regions in 1Q95. The high degree residues involved include Arg-105, Tyr-226 and
 3 Gln-288.

4

5

6 **Table Captions**

7 **Table 1.** The significant amino acids and amino acid groups at various interaction
 8 degrees and structures.

9 **Table 2.** Comparison of 1ML4 and 1Q95 according to the interaction degrees.

10

11

12 **Additional files**

13 **Additional file 1** List of 131 pairs of thermophilic proteins and their mesophilic
 14 homologs

15 **Fig. S1** Comparison of interaction degree and the residue conservation score calculated
 16 with ConSurf (lower values indicate more conserved residues).

17 **Fig. S2** The correlation between ASA and interaction degree.

18

19

20

21

Table 1

Degree/ Structure	Significant AA	Significant AA groups	Degree/ Structure	Significant AA	Significant groups	AA
=1	+E, -S, -T	-AM	=6	-C, -N, -Q, -S, +Y	+AR, -TI, +LA, -SU, -AM	
=2	+E, -Q, +R, - T	+CH, +LA, - AM	=7	-M, +R	-SU	
=3	-C, -Q	-AM	=8	-A, -C, -D, +T, +W, -Y	-TI	
=4	+E, -Q	-SU, -AM	=9	-D, +K, -Q, +R	+CH, +PC	
=5	+E, -Q, +R	+CH, +LA, - AM	≥10	+A, +H, +L, -Q, +R, -W	+CH, +AL, +HY, +PC, +TI, +SM	
Seq.	+E, -Q	-AM	Helix	+E, -Q	-TI, -AM	
Coil	+E	-AM	Sheet	-C, -Q, -S	-TI, -SU, -AM	
Buried	-C	-SU	Exposed	-A, +E, +R, -S	+CH, -NE, +NC, -TI, +LA, -AM	
Partial buried	-Q	-AM	Partial exposed	+E, -Q, +R, -S	+CH, -TI, +LA, -AM	

CH: charged, AL: aliphatic, AR: aromatic, PO: polar, NE: neutral, HY: hydrophobic, PC: positive charged, NC: negative charged, TI: tiny, SM: small, LA: large, SU: sulfur, AM: amide. Seq. sequences. Amino acid is represented as 1-letter. “+” indicates more in thermophilic, “-“ indicates less in thermophilic

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19

Table 2

PDB ID	length	AAs with interactions	Low degree AAs	Medium degree AAs	High degree AAs
1ML4 (T)	308	293 (95.1%)	178 (60.8%)	107(36.5%)	8(2.7%)
1Q95 (M)	310	290 (93.5%)	173 (59.7%)	114(39.3%)	3(1.0%)

1
2
3

4

5 **FIG.1**

6

7

8

9

10

11

12

13

14

FIG.2

FIG.3

FIG.4

1
2
3

4
5

6
7
8
9
10
11
12
13
14
15
16
17
18

FIG.5

FIG.6