IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 19, 2020, accepted April 1, 2020, date of publication April 6, 2020, date of current version April 21, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2985737

The Frequencies of Oppositely Charged, Uncharged Polar, and β-Branched Amino Acids Determine Proteins' Thermostability

SHANWEN SUN[1], CHUNYAN AO[1], DONGHUA WANG[2], AND BENZHI DONG[3]

[1]Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu 610054, China
[2]Department of General Surgery, Heilongjiang Province Land Reclamation Headquarters General Hospital, Harbin 150088, China
[3]Information and Computer Engineering College, Northeast Forestry University, Harbin 150001, China

Corresponding authors: Donghua Wan (wangdonghua7885@163.com) and Benzhi Dong (nefu_dbz@163.com)

This work was supported in part by the National Key R&D Program of China under Grant 2018YFC0910405, and in part by the National Natural Science Foundation of China under Grant 61771331, Grant 61922020, and Grant 91935302.

ABSTRACT Enhancing proteins' thermostability is an important aspect of enzyme engineering. Many studies have investigated the properties that determine the proteins' thermostability. However, no consensus has emerged. To understand the mechanisms underlying the high thermostability of thermophilic proteins, we evaluated the relative importance of the amino acid frequencies in protein sequences for discriminating thermophilic and non-thermophilic proteins based on machine learning algorithms together with a three-step feature selection procedure and a principal component (PC) analysis to remove noisy and redundant information. Our results showed that the frequencies of oppositely charged amino acids, i.e., Lys and Glu, were higher in thermophilic proteins, suggesting that electrostatic interactions are fundamentally important for protein stabilization at high temperatures. Further, we found that the frequencies of uncharged polar amino acids, which are thermolabile or actively interact with water molecules, were lower in thermophilic proteins. Moreover, the frequencies of β-branched aliphatic amino acids tended to increase with increasing thermostability. Overall, these results suggest that proteins' thermostability is determined by a few protein features, which were well captured by the first two PCs. A classifier based on only the first two PCs achieved a high accuracy of 90%, suggesting that our classifier could be an effective and efficient tool for engineering stable proteins.

INDEX TERMS Thermostability, protein, machine learning, amino acid composition.

I. INTRODUCTION

Proteins are important biocatalysts; however, most of them are unstable at high temperatures, severely curtailing their applications in the chemical industry [1]. Many efforts, therefore, have been devoted to enhancing proteins' thermostability. Thermophilic organisms, such as Thermus aquaticus, produce proteins that can tolerate high temperatures even up to 120 °C [2]. These thermophilic proteins are key materials for exploring the mechanisms that allow proteins to maintain stability at high temperatures, and for designing and optimizing enzymes [3].

Some early studies conducted pairwise comparisons of thermophilic and non-thermophilic proteins and found that changes in amino acid residues on the molecular surface can affect proteins' thermostability [4]–[7]. For instance, Argos *et al.* [4] and Haney *et al.* [5] found that replacing certain amino acids, such as Gly and Ser, with Glu can increase the proteins' capacity to tolerate high temperatures. These results suggest that the existence of Glu may increase proteins' thermostability [4], [5]. In contrast, Kawamura *et al.* [6] showed that a replacement of Glu with Gly in the Bacillus stearothermophilus DNA-binding protein [8] HU greatly enhanced the thermostability of the mutant protein. Similarly, Perl *et al.* [7] reported that changing Glu to Arg or Leu could transform a mesophilic protein into a thermophilic protein.

To solve the inconsistency among the results of previous studies, later studies compiled protein sequence data from multiple thermophilic and non-thermophilic organisms and tested the protein features that may

The associate editor coordinating the review of this manuscript and approving it for publication was Jijun Tang.

discriminate thermophilic and non-thermophilic proteins using traditional statistical methods, such as t-tests and linear regressions [9]–[17]. Various protein features were examined in these studies. Fukuchi and Nishikawa [9] reported that the amino acid composition (AAC) of the protein surface was key for the discrimination of thermophilic and non-thermophilic proteins. Das and Gerstein [10] further showed that thermophilic proteins have a higher level of charged residues than non-thermophilic proteins, suggesting that electrostatic interactions, such as ion pairs, play an important role in thermostability. This is in line with several other studies showing that thermophilic proteins had greater polar surface area, more frequent salt bridges and hydrogen bonds, and higher isoelectric points [11], [18]–[22]. Additionally, researchers found a greater number of hydrophobic residues in thermophilic proteins, indicating the importance of hydrophobic interactions for promoting thermostability [11], [20]–[22]. Other features, such as packing density or compactness [23], secondary structural composition [24], changes in entropy upon folding [25], and surface-to-volume ratio [26], were also found to affect proteins' stability at high temperatures. However, the relative importance of various features regarding thermostability remains unknown and is difficult to assess using traditional statistical analyses due to collinearity among features. Another subtle weakness in some of the above studies is that multiple tests were performed without corrections, which may have resulted in an increased type I error [27].

Modern machine learning algorithms are free of statistical tests and assumptions about data distribution and can handle collinearity to some extent. Some algorithms, such as random forest, can also provide an importance assessment for each independent variable. Many studies have attempted to use machine learning algorithms and various features to discriminate thermophilic and non-thermophilic sequences or to predict a given protein's thermostability [28]–[31]. Zhang and Fang [32] tested the performance of AAC to discriminate thermophilic and non-thermophilic sequences based on a support vector machine (SVM), and the accuracy was 0.91 with 5-fold cross-validation. They also compared the influences of four machine learning algorithms on the performance of AAC and found that the algorithms that had the best performance for thermophilic and for non-thermophilic proteins were different, with the best average accuracy being 0.73 [33]. Gromiha and Suresh [28] used AAC and dipeptide composition to predict proteins' thermostability based on 12 algorithms. They found that different algorithms resulted in similar accuracies of around 0.89, and the inclusion of the dipeptide composition did not significantly improve the performance [28]. Similarly, Lin and Chen [29] showed that the addition of key features selected from dipeptide composition only improved the accuracy from 92.56%, predicted solely based on AAC, to 93.27%, suggesting that AAC is the main determinant of proteins' thermostability. Considering that many other features that affect thermostability, such as salt bridges, are also influenced by AAC [25], [34], [35], assessments of the relative importance of different amino acids and the capacity of the combination of key amino acids to distinguish thermophilic and non-thermophilic proteins are, therefore, essential to understanding the mechanisms that determine the high stability of thermophilic proteins.

In this study, we first extracted data on the frequencies of 20 amino acids and then evaluated their importance and selected relevant amino acids that could discriminate thermophilic and non-thermophilic proteins based on a three-step procedure. A dimensionality reduction procedure based on principal component (PC) analysis was used to further remove redundant information among the selected amino acids. The performance of the first two PCs regarding discriminating thermophilic and non-thermophilic proteins was assessed based on jackknife cross-validation for three machine learning algorithms, i.e., SVM, random forest, and regularized logistic regression.

II. MATERIALS AND METHOD

A. DATASETS

The benchmark dataset of thermophilic and non-thermophilic protein sequences was obtained from Lin and Chen [29]. It contained data collected from thermophilic organisms and non-thermophilic organisms in the Universal Protein Resource (UniProt) [29]. Species with a lower limit of optimal growth temperature $\geq$60 °C were defined as thermophilic organisms [29]. In contrast, species with an upper limit of optimal growth temperature $\leq$30 °C were classified as non-thermophilic organisms [29]. Additionally, the sequences were further scrutinized to ensure their reliability, i.e., the sequences were manually annotated and reviewed, were not fragments of other sequences or constructed from prediction or homology, had low sequence similarity (<40%) [36], and had no ambiguous residues [29]. In total, 1329 thermophilic and 1250 non-thermophilic protein sequences from 17 archaea and 119 bacteria were included in the analysis [29].

B. FEATURE EXTRACTION AND SELECTION

AAC was assessed in terms of the frequency of each of the 20 amino acids per protein sequence with the following equation (**Fig. 1**):

$$C_N = \frac{Number\ of\ (N)}{Length\ (protein\ sequence)} \tag{1}$$

where N = Ala, Cys, Asp, Glu, Phe, Gly, His, Ile, Lys, Leu, Met, Asn, Pro, Gln, Arg, Ser, Thr, Val, Trp, and Tyr.

The three-step procedure in the VSURF R package was used to select relevant amino acids for discriminating thermophilic and non-thermophilic proteins based on random forests with ntree = 2000 trees [37]. Specifically, in the first step, a random forest, which was widely employed in bioinformatics [38], [39], was built to assess the variable importance of each amino acid based on the bootstrap samples and the out-of-bag samples, i.e., one-third of the original data that

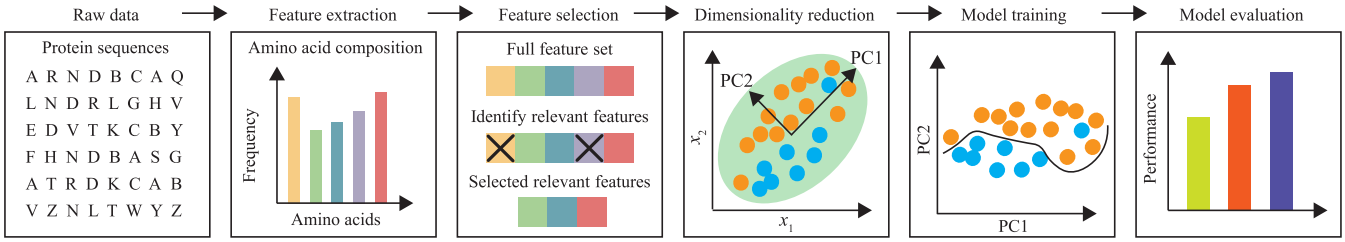


FIGURE 1. Study scheme. Raw protein sequences were first fed into a feature extraction process. The amino acid composition (AAC) was extracted in terms of the frequencies of each amino acid in each protein sequence. Thereafter, a three-step feature selection procedure was used to select amino acids relevant to discriminating thermophilic and non-thermophilic proteins based on random forests. A principal component (PC) analysis was then utilized to further remove redundant information in the selected relevant amino acids. The first two PCs were used to train three classifiers, i.e., support vector machine (SVM), random forest, and regularized logistic regression. Model performances were evaluated using jackknife cross-validation with four evaluation metrics, i.e., sensitivity (SN), specificity (SP), accuracy, and area under the receiver operating characteristic curve (AUC).

were left out of the bootstrap samples [37]. Thereafter, amino acids were ordered according to their variable importance and were eliminated if their importance was smaller than the standard deviation of the variable importance of the useless amino acids [37]. In the second step, nested random forests were constructed. Amino acids retained in step one were successively added into the models based on their variable importance. The amino acids in the model with the smallest out-of-bag error were kept [37]. In step three, amino acids left in the preceding step were listed in descending order based on variable importance and were sequentially included in the random forest models until the decrease of out-of-bag error was smaller or equal to the average variation obtained by introducing noisy variables [37]. The subset of amino acids remaining in the final model was identified as relevant to discriminating thermophilic and non-thermophilic proteins and used for further reducing redundant information (**Fig. 1**).

C. DIMENSIONALITY REDUCTION

PC analysis [40] was utilized to further reduce redundancy among the relevant amino acids (**Fig. 1**). The first two PCs were then used to discriminate thermophilic and non-thermophilic proteins (**Fig. 1**). To assess the importance of each amino acid for the first two PCs, the contribution of each amino acid was assessed based on the following equation:

$$Contribution_i = \frac{cor_i^2}{\sum_i cor_i^2} \times 100 \qquad (2)$$

where cor_i is the correlation coefficient between relevant amino acid i and each PC.

D. DISCRIMINATION OF THERMOPHILIC AND NON-THERMOPHILIC PROTEINS

Three classifiers, SVM [41]–[45], random forest, and regularized logistic regression, were implemented to distinguish thermophilic and non-thermophilic proteins (**Fig. 1**). These models were built and their tuning hyperparameters were optimized using the caret R package [46].

Model performances were evaluated using jackknife cross-validation with four evaluation metrics, i.e., sensitivity (SN), specificity (SP), accuracy, and area under the receiver operating characteristic curve (AUC; **Fig. 1**). SN is the proportion of thermophilic proteins that are correctly predicted as thermophilic proteins. SP is the proportion of non-thermophilic proteins that are correctly predicted as non-thermophilic proteins. Accuracy is the proportion of correctly predicted thermophilic and non-thermophilic proteins. AUC is the area under the curve of SN plotted against (1-SP) and it assesses the model's ability to avoid false prediction.

To illustrate the performance of key amino acids for classifying thermophilic and non-thermophilic proteins, model performances were assessed with three sets of composition matrixes, i.e., all 20 amino acids, the subset of relevant amino acids, and the first two PCs.

III. RESULTS

The amino acid composition differed between thermophilic and non-thermophilic proteins (**Fig. 2 a**). With all 20 amino acids, a high discrimination performance was achieved regardless of the algorithms and evaluation metrics used (**Table 1**). 93% of thermophilic proteins and 93% of non-thermophilic proteins were correctly predicted using SVM and random forest, respectively. The overall accuracy was 91% using either SVM or random forest, and the AUC was 0.98 using SVM and 0.97 using random forest.

The importance of the 20 amino acids varied, with Glu, Gln, and Lys having higher importance than others (**Fig. 2 b**). In total, 14 amino acids were selected as the relevant amino acids for discriminating thermophilic and non-thermophilic proteins (**Fig. 2 b**). With these 14 amino acids, a higher percentage of correctly discriminated proteins (92%) was achieved using random forest than with all 20 amino acids (**Table 1**). The highest AUC and SN were 0.98 and 0.93, respectively, using 14 relevant amino acids, the same as when all 20 amino acids were used, while a slightly lower percentage of non-thermophilic proteins were correctly classified (**Table 1**).

Although the first two PCs only accounted for 33% of the variations in the 14 relevant amino acids, most thermophilic and non-thermophilic proteins were separated along with them (**Fig. 2 c**). The discrimination performance with the first two PCs was comparable to the discrimination performance with all 20 amino acids and with the 14 relevant amino acids (**Table 1**). Using SVM, for instance, 92% of

TABLE 1. Performances of support vector machine (SVM), random forest and logistic regression using three sets of composition matrixes (all 20 amino acids, the subset of relevant amino acids after feature selection, and the first two PCs). Models were evaluated based on jackknife cross-validation.

Composition matrix	Algorithm	SN	SP	Accuracy	AUC
All 20 amino acids	SVM	0.93	0.92	0.91	0.98
	Random forest	0.90	0.93	0.91	0.97
	Logistic regression	0.91	0.90	0.90	0.96
14 relevant amino acids	SVM	0.92	0.91	0.91	0.98
	Random forest	0.93	0.90	0.92	0.97
	Logistic regression	0.90	0.89	0.90	0.96
First two PCs	SVM	0.92	0.88	0.90	0.94
	Random forest	0.90	0.86	0.88	0.94
	Logistic regression	0.90	0.87	0.89	0.95

AUC: area under the receiver operating characteristic curve; PC: principal component; SN: sensitivity; SP: specificity.

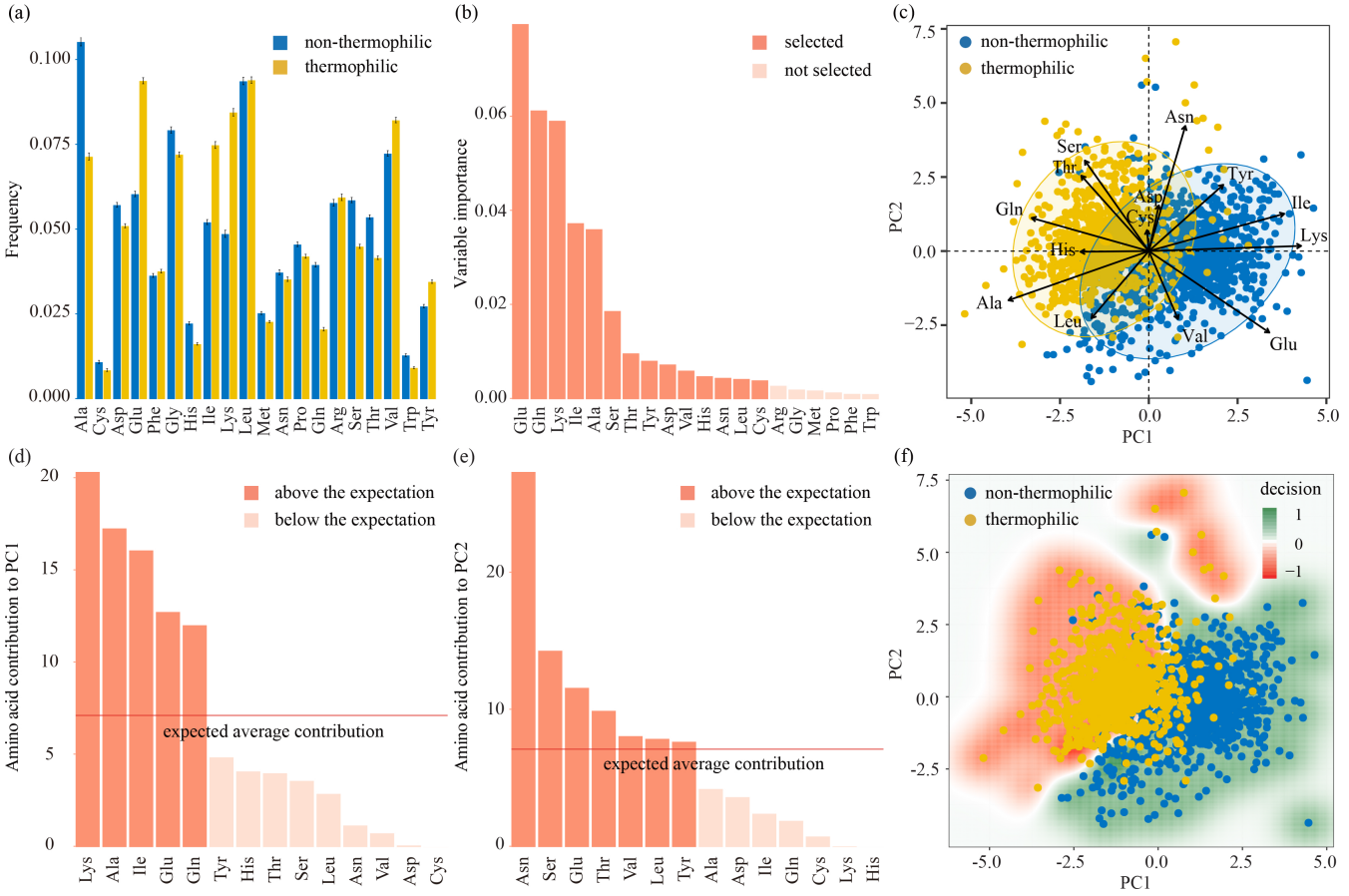


FIGURE 2. (a) The amino acid composition (AAC) in thermophilic and non-thermophilic proteins. The height of each bar indicates the mean frequency of each amino acid in thermophilic proteins or non-thermophilic proteins, and the error bar shows the corresponding standard error. (b) Amino acids ordered according to their importance for discriminating thermophilic and non-thermophilic proteins in the feature selection procedure. Out of the 20 amino acids, 14 relevant and six trivial amino acids were labelled as 'selected' and 'not selected', respectively. (c-e) Results of the principal component (PC) analysis (dimensionality reduction). In (c), the positions of thermophilic and non-thermophilic protein sequences along the first two PCs and the contribution of each selected amino acid to the first two PCs were plotted. The ellipses indicate where 95% of the thermophilic and non-thermophilic proteins were distributed. In (d) and (e), to better visually illustrate the importance of amino acids to each of the first two PCs, the contribution of each amino acid to PC1 and PC2 are presented separately. The red horizontal line in the bar plots indicates the expected average contribution, calculated as 1/14, as 14 relevant amino acids were selected in (b) and the contributions of amino acids were assumed to be uniform [58]. (f) Discrimination of thermophilic and non-thermophilic proteins based on PC1 and PC2 using support vector machine (SVM). The decision values used to discriminate the proteins are also shown.

thermophilic proteins, 88% of non-thermophilic proteins, and 90% of proteins overall were correctly predicted, and the AUC was 0.94 (**Table 1**; **Fig. 2 d**). Logistic regression and random forest performed similarly to SVM (**Table 1**). The results from logistic regression showed that the coefficients of PC1 and PC2 were 0.84 and -0.37, respectively, suggesting that the effect of PC1 on the discrimination of thermophilic proteins was positive while the effect of PC2 was negative.

Among the 14 relevant amino acids, Lys, Ile, Glu, Ala, and Gln had relatively high contributions to PC1 than others, which was generally consistent with their importance according to the feature selection procedure; Ala and Gln were negatively correlated with PC1 (**Fig. 2 d**). Asn, Ser, Thr, Tyr, Glu, Val, and Leu had relatively high contributions to PC2; the latter three were negatively correlated with PC2 (**Fig. 2 e**).

IV. DISCUSSION

We found that a few amino acids determined the proteins' thermostability. After removing the redundant amino acids, 92% of proteins were correctly classified as thermophilic or non-thermophilic proteins (**Table 1**). Our further PCA analyses showed that even based on only the first two PCs that were derived from the 14 relevant amino acids, 90% accuracy was achieved (**Table 1**), suggesting that there was a lot of redundant information in the relevant amino acids and the proteins' thermostability was governed by a small number of amino acid properties.

A. DETERMINANTS OF PROTEINS' THERMOSTABILITY

The largest variation among the frequencies of the 14 selected relevant amino acids was captured by PC1 (**Fig. 2 c**). Along PC1, the predominant trend, which is consistent with their relatively higher variable importance, was the concurrently increased frequencies of Lys and Glu in thermophilic proteins (**Fig. 2 a,** 2 **b-d**). Lys is a positively charged amino acid with an isoelectronic point of 9.74, while Glu is a negatively charged amino acid with an isoelectronic point of 3.22, implying that the existence of oppositely charged amino acids can stabilize proteins at high temperatures. Likewise, many studies have also found a higher level of oppositely charged amino acids in thermophilic proteins than in non-thermophilic proteins [10], [11], [47]. These results suggest that electrostatic interactions are fundamentally important for enhancing proteins' stability at high temperatures [48]. Indeed, salt bridges and hydrogen bonds are crucial for increasing the conformational stability of proteins [14], [49]. The higher number of salt bridges and hydrogen bonds in thermophilic compared to non-thermophilic proteins also supports the role of electrostatic interactions in determining proteins' thermostability [14], [22], [49].

Our results further suggest that a reduction of uncharged polar amino acids in proteins can increase their stability at high temperatures. This is supported by the tendency toward reductions in the frequencies of uncharged polar amino acids, i.e., Asn, Ser, and Thr, in thermophilic proteins along PC2 (**Fig. 2 a,** 2 **c, and** 2 **e**). Consistently, the other uncharged polar amino acid, Gln, also decreased in thermophilic proteins along PC1 (**Fig. 2 a,** 2 **c-d**). Gln and Asn are thermolabile and deamidate at high temperatures [50], [51]. Ser and Thr have relatively more interactions with water molecules surrounding proteins [52] and, at high temperatures, they release water molecules that interacted with them at low temperatures, and thus induce the proteins' instability [35]. Therefore, suitable replacement of uncharged polar amino acids by other amino acids, especially charged ones, leads to fewer thermolabile residues and may lead to improved protein thermostability. Similarly, several studies have reported lower frequencies of uncharged polar amino acids co-occurring with higher frequencies of charged amino acids in thermophilic proteins [14], [47], [53]. Bhanuramanand *et al.* [51] also showed that replacing a few deamidation-susceptible Asn residues resulted in higher thermostability in a lipase. Collectively, these results suggest that the absence of uncharged polar amino acids is important for enhancing proteins' thermostability.

Another noticeable but inconsistent trend is that the frequencies of some aliphatic amino acids, i.e., Ile, Val, and Leu, increased, but the other aliphatic amino acid, Ala, decreased with increasing thermostability (**Fig. 2 a,** 2 **c-e**). Aliphatic amino acids are generally hydrophobic; Ile, Val, and Leu are β-branched amino acids and are the most hydrophobic [54], while Ala is an unbranched amino acid and is less hydrophobic. The strength of hydrophobic interactions formed by these hydrophobic amino acids increases with temperature [55], and may thus be important for maintaining stability at high temperatures. Ikai [56] proposed the aliphatic index, which is mainly determined by the frequencies of Ile, Val, and Leu in protein sequences, and reported that the aliphatic index positively contributed to proteins' thermostability. Likewise, Lu *et al.* [57] found that thermophilic proteins had a higher aliphatic index and higher Leu composition than non-thermophilic proteins. High levels of β-branched amino acids rather than unbranched amino acids can also lead to a smaller increase in conformational entropy upon unfolding [53]. The opposite correlations of Ile and Ala with PC1 thus suggest that a high frequency of Ala may be replaced by a high frequency of Ile in thermophilic proteins to increase their thermostability. Similarly, Chakravarty and Varadarajan [53] found that the frequency of Ile was slightly higher and the frequency of Ala was slightly lower in thermophilic proteins than in non-thermophilic proteins, although the differences were not significant. In contrast, by conducting pairwise comparisons of thermophilic and non-thermophilic proteins, Argos *et al.* [4] showed that the high frequency of Val was replaced by Ile and Ala in thermophilic proteins, but the latter replacement may decrease internal hydrophobicity and packing. Nevertheless, our results are in line with the higher levels of Ile, Val, and Leu found in thermophilic proteins [14], [53], [57], indicating the importance of β-branched amino acids for increasing proteins' thermostability.

B. DISCRIMINATION OF THERMOPHILIC AND NON-THERMOPHILIC PROTEINS

Two things were found to be important for building classifiers to distinguish thermophilic and non-thermophilic proteins in our study. The first is to use AAC, in terms of the frequencies of amino acids. Based on the frequencies of each of the 20 amino acids, 91% of proteins were correctly discriminated in our study (**Table 1**). Similarly, Zhang and Fang [32] and Gromiha and Suresh [28] used the frequencies of all 20 amino acids to predict a protein's thermostability,

and the accuracy was 90.5% and 89%, respectively. In contrast, Miotto *et al.* [31] used proteins' secondary structure content and architecture to discriminate thermophilic and non-thermophilic proteins, but only 76% of proteins were correctly discriminated. These results suggest that AAC, rather than other metrics, appropriately captures the mechanisms underlying proteins' thermostability, i.e., the frequencies of oppositely charged, uncharged polar, and β-branched amino acids in a protein.

Further, we found that using feature selection procedures to eliminate noisy and irrelevant amino acids improved accuracy. In our study, six amino acids were assessed as trivial based on a three-step feature selection procedure, and after removing these amino acids, the accuracy improved up to 92% (**Table 1**). In accordance with our result, Lin and Chen [29] employed an analysis of variance (ANOVA) feature selection procedure to select ten relevant dipeptides, and by combining them with the frequencies of all 20 amino acids, they achieved an accuracy of 93.3%. Wang and Li [30] further improved the accuracy to 95% by selecting nine amino acids and 38 dipeptides using a genetic algorithm. These results thus highlight the necessity of performing a feature selection procedure for building classifiers to distinguish thermophilic and non-thermophilic proteins.

V. CONCLUSION

Understanding the mechanisms that determine proteins' thermostability and building classifiers to discriminate thermophilic and non-thermophilic proteins are important for engineering enzymes that are stable at high temperatures. We found that proteins' thermostability is determined by a few protein features. First, thermostability was increased mainly due to stronger electrostatic interactions formed by oppositely charged amino acids (especially Lys and Glu). Additionally, thermostability was increased by decreasing the number of uncharged polar amino acids, which minimized deamidation at high temperatures and reduced the interactions with water molecules. Moreover, the number of β-branched amino acids can also affect thermostability. Overall, these protein features were appropriately captured by the first two PCs derived from the frequencies of relevant amino acids that were selected based on a feature selection procedure. A high discrimination accuracy of 90% was achieved with only PC1 and PC2 in our study. Therefore, our classifier is effective and efficient at discriminating thermophilic and non-thermophilic proteins and useful for designing thermostable proteins.

REFERENCES

[1] A. S. Bommarius, J. M. Broering, J. F. Chaparro-Riggers, and K. M. Polizzi, "High-throughput screening for enhanced protein stability," *Current Opinion Biotechnol.*, vol. 17, no. 6, pp. 606–610, Dec. 2006.

[2] S.-L. Huang, L.-C. Wu, H.-K. Liang, K.-T. Pan, J.-T. Horng, and M.-T. Ko, "PGTdb: A database providing growth temperatures of prokaryotes," *Bioinformatics*, vol. 20, no. 2, pp. 276–278, Jan. 2004.

[3] H. Tang, R.-Z. Cao, W. Wang, T.-S. Liu, L.-M. Wang, and C.-M. He, "A two-step discriminated method to identify thermophilic proteins," *Int. J. Biomath.*, vol. 10, no. 04, May 2017, Art. no. 1750050.

[4] P. Argos, M. G. Rossman, U. M. Grau, H. Zuber, G. Frank, and J. D. Tratschin, "Thermal stability and protein structure," *Biochemistry*, vol. 18, no. 25, pp. 5698–5703, Dec. 1979.

[5] P. J. Haney, J. H. Badger, G. L. Buldak, C. I. Reich, C. R. Woese, and G. J. Olsen, "Thermal adaptation analyzed by comparison of protein sequences from mesophilic and extremely thermophilic Methanococcus species," *Proc. Nat. Acad. Sci. USA*, vol. 96, no. 7, pp. 3578–3583, Mar. 1999.

[6] S. Kawamura, Y. Kakuta, I. Tanaka, K. Hikichi, S. Kuhara, N. Yamasaki, and M. Kimura, "Glycine-15 in the bend between two ±-Helices can explain the thermostability of DNA binding protein HU fromBacillus stearothermophilus," *Biochemistry*, vol. 35, no. 4, pp. 1195–1200, Jan. 1996.

[7] D. Perl, U. Mueller, U. Heinemann, and F. X. Schmid, "Two exposed amino acid residues confer thermostability on a cold shock protein," *Nat. Struct. Biol.*, vol. 7, no. 5, pp. 380–383, May 2000.

[8] K. Qu, L. Wei, and Q. Zou, "A review of DNA-binding proteins prediction methods," *Current Bioinf.*, vol. 14, no. 3, pp. 246–254, Mar. 2019.

[9] S. Fukuchi and K. Nishikawa, "Protein surface amino acid compositions distinctively differ between thermophilic and mesophilic bacteria," *J. Mol. Biol.*, vol. 309, no. 4, pp. 835–843, Jun. 2001.

[10] R. Das and M. Gerstein, "The stability of thermophilic proteins: A study based on comprehensive genome comparison," *Funct. Integrative Genomics*, vol. 1, no. 1, pp. 76–88, May 2000.

[11] T. J. Taylor and I. I. Vaisman, "Discrimination of thermophilic and mesophilic proteins," *BMC Struct. Biol.*, vol. 10, p. S5, Aug. 2010.

[12] M. M. Gromiha, M. Oobatake, and A. Sarai, "Important amino acid properties for enhanced thermostability from mesophilic to thermophilic proteins," *Biophys. Chem.*, vol. 82, no. 1, pp. 51–67, Nov. 1999.

[13] J. Hasegawa, "Selected mutations in a mesophilic cytochrome c confer the stability of a thermophilic counterpart," *J. Biol. Chem.*, vol. 275, no. 48, pp. 37824–37828, Dec. 2000.

[14] S. Kumar, C.-J. Tsai, and R. Nussinov, "Factors enhancing protein thermostability," *Protein Eng., Des. Selection*, vol. 13, no. 3, pp. 179–191, Mar. 2000.

[15] A. S. Panja, B. Bandopadhyay, and S. Maiti, "Protein thermostability is owing to their preferences to non-polar smaller volume amino acids, variations in residual physico-chemical properties and more salt-bridges," *PLoS ONE*, vol. 10, no. 7, 2015, Art. no. e0131495.

[16] T. J. Taylor and Vaisman, II, "Discrimination and classification of thermophilic and mesophilic proteins," in *4th Int. Symp. Voronoi Diagrams Sci. Eng. (ISVD)*, Glamorgan, U.K., 2007, pp. 212–221.

[17] N. Kannan and S. Vishveshwara, "Aromatic clusters: A determinant of thermal stability of thermophilic proteins," *Protein Eng., Des. Selection*, vol. 13, no. 11, pp. 753–761, Nov. 2000.

[18] T. Kawashima, N. Amano, H. Koike, S.-I. Makino, S. Higuchi, Y. Kawashima-Ohya, K. Watanabe, M. Yamazaki, K. Kanehori, T. Kawamoto, T. Nunoshiba, Y. Yamamoto, H. Aramaki, K. Makino, and M. Suzuki, "Archaeal adaptation to higher temperatures revealed by genomic sequence of thermoplasma volcanium," *Proc. Nat. Acad. Sci. USA*, vol. 97, no. 26, pp. 14257–14262, Dec. 2000.

[19] G. Vogt, S. Woell, and P. Argos, "Protein thermal stability, hydrogen bonds, and ion pairs," *J. Mol. Biol.*, vol. 269, no. 4, pp. 631–643, Jun. 1997.

[20] E. Christodoulou, W. R. Rypniewski, and C. E. Vorgias, "High-resolution X-ray structure of the DNA-binding protein HU from the hyperthermophilic thermotoga maritima and the determinants of its thermostability," *Extremophiles*, vol. 7, no. 2, pp. 111–122, Apr. 2003.

[21] M. M. Gromiha, M. C. Pathak, K. Saraboji, E. A. Ortlund, and E. A. Gaucher, "Hydrophobic environment is a key factor for the stability of thermophilic proteins," *Proteins, Struct., Function, Bioinf.*, vol. 81, no. 4, pp. 715–721, Apr. 2013.

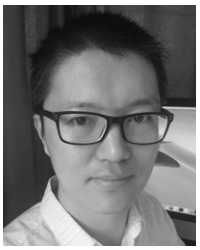
[22] M. Sadeghi, H. Naderi-Manesh, M. Zarrabi, and B. Ranjbar, "Effective factors in thermostability of thermophilic proteins," *Biophysical Chem.*, vol. 119, no. 3, pp. 256–270, Feb. 2006.

[23] R. J. Russell, D. W. Hough, M. J. Danson, and G. L. Taylor, "The crystal structure of citrate synthase from the thermophilic archaeon, thermoplasma acidophilum," *Structure*, vol. 2, no. 12, pp. 1157–1167, Dec. 1994.

[24] A. Szilágyi and P. Závodszky, "Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits: Results of a comprehensive survey," *Structure*, vol. 8, no. 5, pp. 493–504, May 2000.

[25] L. Sawle and K. Ghosh, "How do thermophilic proteins and proteomes withstand high temperature?" *Biophysical J.*, vol. 101, no. 1, pp. 217–227, Jul. 2011.

[26] J. J. Tanner, R. M. Hecht, and K. L. Krause, "Determinants of enzyme thermostability observed in the molecular structure of thermus aquaticus d-glyceraldehyde-3-phosphate dehydrogenase at 2.5 resolution," *Biochemistry*, vol. 35, no. 8, pp. 2597–2609, 1996.
[27] W. S. Noble, "How does multiple testing correction work?" *Nature Biotechnol.*, vol. 27, no. 12, pp. 1135–1137, Dec. 2009.
[28] M. M. Gromiha and M. X. Suresh, "Discrimination of mesophilic and thermophilic proteins using machine learning algorithms," *Proteins, Struct., Function, Bioinf.*, vol. 70, no. 4, pp. 1274–1279, 2008.
[29] H. Lin and W. Chen, "Prediction of thermophilic proteins using feature selection technique," *J. Microbiological Methods*, vol. 84, no. 1, pp. 67–70, Jan. 2011.
[30] L. Wang and C. Li, "Optimal subset selection of primary sequence features using the genetic algorithm for thermophilic proteins identification," *Biotechnol. Lett.*, vol. 36, no. 10, pp. 1963–1969, Oct. 2014.
[31] M. Miotto, P. P. Olimpieri, L. Di Rienzo, F. Ambrosetti, P. Corsi, R. Lepore, G. G. Tartaglia, and E. Milanetti, "Insights on protein thermal stability: A graph representation of molecular interactions," *Bioinformatics*, vol. 35, no. 15, pp. 2569–2577, Aug. 2019.
[32] G. Zhang and B. Fang, "Support vector machine for discrimination of thermophilic and mesophilic proteins based on amino acid composition," *Protein Peptide Lett.*, vol. 13, no. 10, pp. 965–970, Oct. 2006.
[33] G. Zhang and B. Fang, "Discrimination of thermophilic and mesophilic proteins via pattern recognition methods," *Process Biochem.*, vol. 41, no. 3, pp. 552–556, Mar. 2006.
[34] E. Christodoulou and C. Vorgias, "The thermostability of DNA-binding protein HU from mesophilic, thermophilic, and extreme thermophilic bacteria," *Extremophiles*, vol. 6, no. 1, pp. 21–31, Feb. 2002.
[35] X.-X. Zhou, Y.-B. Wang, Y.-J. Pan, and W.-F. Li, "Differences in amino acids composition and coupling patterns between mesophilic and thermophilic proteins," *Amino Acids*, vol. 34, no. 1, pp. 25–33, Jan. 2008.
[36] Q. Zou, G. Lin, X. Jiang, X. Liu, and X. Zeng, "Sequence clustering in bioinformatics: An empirical study," *Briefings Bioinf.*, vol. 21, pp. 1–10, Sep. 2018.
[37] R. Genuer, J.-M. Poggi, and C. Tuleau-Malot, "VSURF: An r package for variable selection using random forests," *Radio J.*, vol. 7, no. 2, pp. 19–33, 2015.
[38] Z. Lv, S. Jin, H. Ding, and Q. Zou, "A random forest sub-golgi protein classifier optimized via dipeptide and amino acid composition features," *Frontiers Bioeng. Biotechnol.*, vol. 7, pp. 1–11, Sep. 2019.
[39] X. Ru, L. Li, and Q. Zou, "Incorporating distance-based Top-n-gram and random forest to identify electron transport proteins," *J. Proteome Res.*, vol. 18, no. 7, pp. 2931–2939, Jul. 2019.
[40] C. Jia, Y. Zuo, and Q. Zou, "O-GlcNAcPRED-II: An integrated classification algorithm for identifying O-GlcNAcylation sites based on fuzzy undersampling and a K-means PCA oversampling technique," *Bioinformatics*, vol. 34, no. 12, pp. 2029–2036, Jun. 2018.
[41] L. Chao, L. Wei, and Q. Zou, "SecProMTB: A SVM-based Classifier for Secretory Proteins of Mycobacterium tuberculosis with Imbalanced Data Set," *Proteomics*, vol. 19, pp. 1–8, Aug. 2019.
[42] H. Bu, J. Hao, J. Guan, and S. Zhou, "Predicting enhancers from multiple cell lines and tissues across different developmental stages based on SVM method," *Current Bioinf.*, vol. 13, no. 6, pp. 655–660, Nov. 2018.
[43] C. Meng, S. Jin, L. Wang, F. Guo, and Q. Zou, "AOPs-SVM: A sequence-based classifier of antioxidant proteins using a support vector machine," *Frontiers Bioeng. Biotechnol.*, vol. 7, pp. 1–10, Sep. 2019.
[44] N. Zhang, Y. Sa, Y. Guo, W. Lin, P. Wang, and Y. Feng, "Discriminating ramos and jurkat cells with image textures from diffraction imaging flow cytometry based on a support vector machine," *Current Bioinf.*, vol. 13, no. 1, pp. 50–56, Feb. 2018.
[45] Y. Wang, F. Shi, L. Cao, N. Dey, Q. Wu, A. S. Ashour, R. S. Sherratt, V. Rajinikanth, and L. Wu, "Morphological segmentation analysis and texture-based support vector machines classification on mice liver fibrosis microscopic images," *Current Bioinf.*, vol. 14, no. 4, pp. 282–294, Apr. 2019.
[46] M. Kuhn, "Building Predictive Models in R Using the caret Package," *J. Stat. Softw.*, vol. 28, no. 5, pp. 1–26, Nov. 2008.
[47] S. P. Pack and Y. J. Yoo, "Protein thermostability: Structure-based difference of amino acid between thermophilic and mesophilic proteins," *J. Biotechnol.*, vol. 111, no. 3, pp. 269–277, Aug. 2004.
[48] L. Xiao and B. Honig, "Electrostatic contributions to the stability of hyperthermophilic proteins," *J. Mol. Biol.*, vol. 289, no. 5, pp. 1435–1444, 1999.
[49] G. Vogt and P. Argos, "Protein thermal stability: Hydrogen bonds or internal packing?" *Folding Des.*, vol. 2, pp. S40–S46, Jun. 1997.
[50] F. Catanzano, G. Barone, G. Graziano, and S. Capasso, "Thermodynamic analysis of the effect of selective monodeamidation at asparagine 67 in ribonuclease a," *Protein Sci.*, vol. 6, no. 8, pp. 1682–1693, Aug. 1997.
[51] K. Bhanuramanand, S. Ahmad, and N. M. Rao, "Engineering deamidation-susceptible asparagines leads to improved stability to thermal cycling in a lipase," *Protein Sci.*, vol. 23, no. 10, pp. 1479–1490, Oct. 2014.
[52] J. M. Goodfellow, N. Thanki, and J. M. Thornton, "Preliminary analysis of water molecule distributions in proteins," *Mol. Simul.*, vol. 3, nos. 1–3, pp. 167–182, May 1989.
[53] S. Chakravarty and R. Varadarajan, "Elucidation of determinants of protein stability through genome sequence analysis," *FEBS Lett.*, vol. 470, no. 1, pp. 65–69, Mar. 2000.
[54] T. E. Creighton, *Proteins: Structures and Molecular Properties*. New York, NY, USA: W. H. Freeman, 1993, pp. 6–19.
[55] R. L. Baldwin, "Temperature dependence of the hydrophobic interaction in protein folding.," *Proc. Nat. Acad. Sci. USA*, vol. 83, no. 21, pp. 8069–8072, Nov. 1986.
[56] A. Ikai, "Thermostability and aliphatic index of globular proteins," *J. Biochem.*, vol. 88, no. 6, pp. 1895–1898, 1980.
[57] B. Lu, G. Wang, and P. Huang, "A comparison of amino acid composition of proteins from thermophiles and mesophiles," *Wei Sheng Wu Xue Bao*, vol. 38, no. 1, pp. 20–25, Feb. 1998.
[58] (2017). *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*. [Online]. Available: https://CRAN.R-project.org/package=factoextra



SHANWEN SUN received the Ph.D. degree from the University of Bayreuth, Germany, in 2019. He is currently a Postdoctoral Researcher with the Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, China. His research interests include bioinformatics and machine learning.

CHUNYAN AO received the M.S. degree from the College of Chemistry, Sichuan University, Sichuan, China, in 2016. Her research interests are bioinformatics and machine learning.

DONGHUA WANG received the bachelor's and master's degrees in medicine from the Harbin Medical School, in 1989 and 1992, respectively. From 2004 to 2019, he was a Chief Physician with the General Hospital of the Provincial Agricultural Reclamation. He is currently the Vice Chairman of the Association, the Executive Director of the Provincial Medical Association, the Executive Director of the Provincial Association of doctors, and the Executive Director of the Association of Youth Prosperity Leaders of the Nongnongken. His research is mainly focused on general surgery. He is a General Member of the Provincial Medical Association and a member of the Cancer Committee of the Provincial Medical Association.

BENZHI DONG was born in Heilongjiang, China, in 1975. He received the B.S. degree in computer engineering from Shenyang Ligong University, in 1997, the M.S. degree in computer science and technology from the Harbin Institute of Technology, in 2004, and the Ph.D. degree in mechanical design and theory from Northeast Forestry University, in 2010. From 1997 to 2008, he was a Lecturer with the College of Computer Science and Engineering, Northeast Forestry University, Harbin, China, where he has been an Assistant Professor with the College of Computer Science and Engineering, since 2008. He is the author of more than 40 articles. His research interests include CAD/CAM, computer vision, and computer recognition of plants and insects. He was a member of China Computer Society.