Bioinformatics Approach to Classification of Four Classes of Organism in Relation to Their Optimal Growth Temperature

Hanaa M. Hussain
The Public Authority of Applied Education and Training, College of Technological Studies, Department of Electronics Engineering Technology, Shuwaikh, Kuwait
Email: hmh.hussain@paeet.edu.kw

Huseyin Seker and Malde Gorania
The University of Northumbria, Department of Computer Sciences and Digital Technology, Faculty of Engineering and Environment Newcastle, Newcastle Upon-Tyne, United Kingdom
Email: huseyin.seker@northumbria.ac.uk

***Abstract*—Identifying the temperature class of proteins in prokaryotic organisms is one of the vital problems in enzyme and protein engineering. In this work, an efficient K-NN predictive models have been developed to discriminate hyperthermophilic, thermophilic, psychrophilic, and mesophilic proteins using Amino acid and Pseudo amino acid compositions. The two predictive models were built and tested with a large dataset consisting of 6631 hyperthermophiles, 11,700 thermophiles, 6267 psychrophiles, and 67,037 mesophiles. Implementation and analysis results showed that the proposed K-NN based predictive models were capable of discriminating the four classes efficiently and with high accuracies, whereby the Amino acid composition model achieved 94% accuracy when using 10-fold cross-validation, and 98% when using hold-out test. on the other hand, the Pseud amino acid composition based model achieved an accuracy of 99% using hold-out test.**

***Index Terms*—amino acid composition, data mining, k-nearest neighbors, machine learning, optimal growth temperature, predictive model, proteins, proteomics, pseudo amino acid composition, thermostability**

I. Introduction

Rapid growth in omics data due to advanced data collection technologies as well as genome sequencing is responsible for the exponential growth in genomics and proteomics databases. It has been estimated that the amount of sequence data has been doubling every seven months over the last decade [1]. In addition, current advancements in biotechnology and data storage have been responsible for the accumulation of large amount of biological data. Consequently, the development of intelligent computational methods is very crucial today to keep up with high computational demands to analyze big data, to extract new knowledge from them, and to identify significant patterns that may help in disease treatment or drug discovery [2]-[4]. Protein and enzyme engineering research has been focusing on identifying features of protein's sequences of some prokaryotes that enable them to grow or function at extreme high or low temperatures. There are multiple factors that affect protein thermostability some of which are still uncovered [5]. Therefore, identifying these factors has been an interesting area of research nowadays due to the vast amount of applications that can benefit from such knowledge including drug design, enzyme and protein engineering [6]. This research intends to investigate the discriminative power of some features of proteins in four classes of prokaryotes and learn whether the K-NN algorithm to efficiently identify these groups.

Prokaryote's are divided into four temperature classes depending on their optimal growth temperature (OPT). These classes are hyperthermophiles, whose OPT > 80° C, thermophiles with OPT of 45-80 ° C, mesophiles whose OPT 20-45 ° C, and psychrophiles with OPT below 20 ° C [7]. This ability to tolerate very high or low temperatures are linked to the protein sequence of each of these four classes, whereby some researchers have found that some particular amino acids were higher in thermophilic proteins than in mesophiles [7]-[10]. Other characteristics of prokaryote's proteins which affect their thermostability include hydrogen bonds, salt bridges, ion pairs, charged residues, and hydrophilic interactions [11]-[13].

A number of research articles proposed methods to discriminate mesophilic and thermophilic proteins based on their amino acid compositions (ACC's) and other physio-chemical properties of proteins using machine learning algorithms. Zhang and Fang used back-check test to discriminate between mesophilic and thermophilic proteins using 400 dipeptide compositions and 20 AAC's [10], [11], and [14]. They analysed a dataset of 8416 proteins and obtained a five-fold cross-validation accuracy of 86.3%. Subsequently, Gromiha and Suresh [9]

Manuscript received May 24 2018; revised July 15, 2018

doi: 10.18178/ijpmbs.7.4.78-83

applied several machine learning algorithms to discriminate the two classes using AAC features of 4684 proteins. They found that the five-fold cross-validation accuracy is almost similar in all the algorithms they used except with neural networks which was slightly higher than the others [9]. In 2011, Lin and Chen [15] constructed a new dataset based on ACC features using 1708 proteins from thermophiles and mesophiles and applied several data mining algorithms to discriminate two classes. The highest achieved performance was for the SVM algorithm, which achieved an accuracy of 93.3% using jackknife cross-validation [15]. In 2012, Nakariyakul et al. [16] used an improved forward floating selection (IFFS) algorithm to reduce the ACC and dipeptide compositions from 420 features to 28 features only, then they applied SVM to discriminate thermophiles and mesophiles using the same datasets used in [15]. The jackknife cross-validation accuracy was 93.3% which was similar to the accuracy achieved in [15] except that they used much less features. In 2013, Zuo et al. [17] developed a robust K-NN-ID classifier, which achieved a jackknife cross-validation accuracy of 91 % using the same two datasets used in [8], [9], and [11]. In 2014, Wang and Li [18] used genetic algorithm coupled with multiple linear regression (MLR) to extract features from ACC and g-gap dipeptide compositions. They were able to extract 9 ACC features, 38 0-gap, and 29 1-gap features. The Jackknife cross-validation achieved an accuracy of 95.4% to discriminate thermophilic proteins from non-thermophilic ones [18]. More recently, Fan et al. [19] developed SVM prediction model based on combining the ACC features, evolutionary information and acid dissociation constant (PKa) to identify thermophilic proteins, which resulted in a dataset of 460 features. The obtained Jackknife cross-validation accuracy was 93.53%, using the same datasets that were used in [9]. More researchers have developed other prediction models to discriminate between the two classes' [20]-[23]. Although many articles have reported high accuracies to discriminate the two classes, there is a necessity to improve the accuracies of prediction furthermore and investigate the predictive power of other physio-chemical characteristics of proteins.

In this research, the authors propose the use of K-nearest neighbors (K-NN) classifier to predict the four protein classes, namely, hyperthermophiles, thermophiles, psychrophiles, and mesophiles by using a newly constructed large dataset utilizing ACC features as well as Pseudo amino acid compositions (PAAC). This study is the first of its kind to our knowledge which tries to discriminate the four classes, and uses a dataset of such large size. In addition, this research will include psychrophiles, which has less been explored in the literature compared with thermophiles and mesophiles.

II. Material and Methods

A. Description of the Datasets

First, large datasets of protein sequences from the four temperature classes were collected from two online databases over the last few years. The first database was the Prokaryotic Temperature Database (PGTdb), which holds a total of 1334 growth temperature from 1072 prokaryotic organisms, namely bacteria and archaea [24]. The second database was the Protein databank (PDB), which is usually used to extract structural information as well as amino acid sequences [25]. Other databases which can be used to extract information about protein thermostability are UniProt, Swiss-Prot, ProTherm [26], and ProtDataTherm [27]. The final datasets which were used in forming the predictive models consists of 6631 hyperthermophiles, 11700 thermophiles, 6267 psychrophiles, and 67037 mesophiles. These protein sequences have been used to develop and validate the K-NN predictive models.

Second, the collected protein sequences were used to compute physio-chemical and structural properties of the proteins that were used in building and testing the predictive models. In this work, ProFeat web server was used to compute protein sequence features, which included (ACC) and PACC [28]. ProFeat, is a web server tool for computing over 1500 commonly used physio-chemical and structural features of proteins from their protein sequences.

Then, the data were checked for redundancy, missing data were removed, and the data were normalized.

Finally, two main K-NN based predictive models were developed one is using the 20 ACC features while the other using the 50 PACC features. PACC features sometimes referred to as the Chou's general Pseudo amino acid composition, named after Kuo Chen Chou who first introduced them in 2001. Since then, the concept of PACC has been adapted and used by many researchers in computational proteomics [2] and [4].

B. The K-Nearest Neighbor (K-NN) Classifier

The K-NN classifier is non-parametric, simple, and fast supervised machine learning method which has been used in mining genomics and proteomics data for class discovery, to define new un-recognized cancer subtypes or classes [30]. Additionally, K-NN has been used in the prediction of protein structural and thermal classes [17] and [31], protein β-turn prediction [32], prediction of protein secondary structure [33, 34], and prediction of melting point of several molecules [35].

The K-NN classifier assign a query vector to the class of its nearest neighbor by computing the distances between the query and the training features using a particular distance metric such as Euclidean, City Block, Cosine, or any others. Depending on the selected "K" values, the resulted distances are then compared and the classifier performs a majority voting to assign the query to the class label of the most encountered class. Several distance metrics may be selected such as Euclidean, Manhattan, Chebychev, Hamming, or others. In this work, Euclidean distance is used, and the value of "K" is varied between 1 to 21 (odd numbers only). The Euclidean distance between the two points x and y is computed as in (1):

$$\text{Euclidean Distance} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}. \qquad (1)$$

C. Implementation Tool

Matlab R2016b, Matlab machine learning and statistical toolboxes were used to implement, validate, and test the K-NN classifier. The hardware used was a workstation running an intel core i7-7820HQ CPU @ 2.9 GHz, and 16 GB RAM.

D. Performance Evaluation Method

In this work, two validation methods were used to assess the performance of the predictive models, one is the 10-fold cross- validation and the other is a hold-out test. In the later, the datasets were divided such that 70% were used for building the model while the remaining 30% were used for validation. This method of validation was selected because it is more suitable for the analysis of large dataset.

Each model was tested 10 times for each value of "K" ranging from 1 to 21, taking odd numbers only. The results from each run were then averaged and the standard deviations were computed.

The performance of each run was assessed by these measures: overall accuracy, specific class accuracy, sensitivity, and specificity, which are usually used in assessing supervised machine learning algorithms [4], equations (2)-(5) illustrates how these assessment measures were computed:

$$\text{Overall Accuracy} = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (2)$$

$$\text{CLASS ACCURACY} = \frac{(TP)\text{ OF THE CLASS}}{\text{NUMBER OF PROTEIN IN THE CLASS}} \quad (3)$$

$$\text{SENSITIVITY} = \frac{TP}{(TP+FN)} \quad (4)$$

$$\text{SPECIFICITY} = \frac{TN}{(TN+FP)} \quad (5)$$

where, TP, FP, TN, and FN refer to true positives (positive instance predicted as positive), false positives (negative instance predicted as positive), true negatives (negative instance predicted as negative), and false negative (positive instance predicted as negative), respectively. These values were extracted from the confusion matrices for each run. The accuracy of the model is a measure of the ability of the model to predict a class label correctly, while the sensitivity and specificity of the predictive model reflects the true positive and true negative rates, respectively.

III. Results and Discussion

The first predictive model which was based on using the 20 ACC features was validated with two methods, one is the 10-fold cross-validation, while the other is the 30% hold-out test. The details of the fours class dataset is shown in Table I, where a total of 64,306 proteins were used for training the model while 27,559 were used for testing them for the case of the 30% hold-out validation. On the other hand, the whole 91,635 proteins were used when performing 10-fold cross-validation. The second predictive model which was based on using PACC, 50 features were used. To validate this model, only 30% hold-out test was used.

TABLE I. Distribution of the Four Class Dataset

Class	Training	Test	Total
Hypethermophiles (C1)	4542	1989	6631
Thermophiles (C2)	8351	3578	11,700
Psychrophiles (C3)	4387	1880	6267
Mesophiles (C4)	46,926	20,111	67,037
Total	**64,306**	**27,559**	**91,635**

A. Classification of the Four Protein Classes Based on ACC Features

The results obtained for the K-NN classification for different values of "K" are presented in Table II and Table III based on 10-fold cross-validation and 30% hold-out test, respectively. These results show that the K-NN classifier can effectively discriminate the four classes with high accuracies. However, when the values of "K" were increased, the sensitivity of the models dropped down. The highest achieved accuracies for the case of K=1 in both models were 93.5%, and 97.8%, respectively, whereas the specificities were 98%, 99%, respectively. On the other hand, the obtained sensitivities were 76% and 91%, respectively.

TABLE II. Performance Results of the ACC K-NN Predictive Model Using 10-Fold Cross-Validation Test

K	C1 %	C2 %	C3 %	C4 %	Accuracy %	Specificity %	Sensitivity %
1	75.02±0.27	73.94 ±0.28	78.23 ±0.32	100	93.53 ±0.22	98.15 ±.11	75.87 ±1.28
3	73.98 ±0.40	72.59 ±0.57	72.42 ±2.15	100	92.8 ±0.33	97.91 ±.13	74.74 ±1.54
5	73.25 ±0.20	72.93 ±0.19	70.22 ±0.23	100	92.50 ±0.26	97.91 ±0.19	73.41 ±1.73
7	72.13 ±0.24	72.49 ±0.10	67.46 ±0.25	100	92.19 ±0.18	97.91 ±0.14	72.08 ±1.12
9	71.21 ±0.21	72.73 ±0.16	65.62 ±0.25	100	92.10 ±0.14	97.97 ±0.23	71.52 ±2.08
11	70.87 ±0.34	72.80 ±0.19	64.69 ±0.35	100	91.8 ±0.26	97.89 ±0.15	70.38 ±1.30
13	70.57 ±0.34	72.95 ±0.28	63.59 ±0.18	100	91.73 ±0.23	97.87 ±0.15	70.14 ±1.90
15	70.29 ±0.29	73.26 ±0.19	62.8 ±0.21	100	91.89 ±0.13	97.93 ±0.16	70.37 ±1.59
17	69.57 ±0.29	73.21 ±0.17	62.4 ±0.24	100	91.85 ±0.20	97.99 ±0.16	69.38 ±1.98
19	68.92 ±0.29	73.18 ±0.22	61.59 ±0.33	100	91.69 ±0.13	97.96 ±0.15	68.96 ±1.98
21	68.42 ±0.26	73.33 ±0.16	60.7 ± 0.36	100	91.54 ±0.24	97.94 ±0.13	68.63 ±2.86

where, C1= Hyperthermophilic, C2= Thermophilic, C3= Psychrophilic, C4= Meosphilic, C1, C2, C3, & C4 are the mean accuracies for each class and corresponding standard deviation.

TABLE III. PERFORMANCE RESULTS OF THE ACC K-NN PREDICTIVE MODEL USING 70% TRAINING AND 30% TEST

K	C1 %	C2 %	C3 %	C4 %	Accuracy %	Specificity %	Sensitivity %
1	91.41 ±0.65	91.38 ±0.64	93.67 ±0.62	100	97.84 ±0.09	99.03 ±0.09	91.40 ±0.66
3	79.71 ±0.54	80.88 ±0.49	83.38 ±0.79	100	94.92 ±0.07	97.59 ±0.7	79.71 ±0.54
5	76.57 ±0.61	78.30 ±0.44	79.31 ±0.74	100	94.08 ±0.09	97.27 ±0.06	76.57 ±0.61
7	73.91 ±0.70	76.86 ±0.78	77.05 ±0.77	100	93.54 ±0.11	97.19 ±0.12	73.91 ±0.69
9	72.17 ±0.76	75.46 ±0.38	73.90 ±0.97	100	93.02 ±0.09	96.97 ±.05	72.17 ±0.77
11	71.59 ±0.76	75.66 ±0.55	71.78 ±0.62	100	92.86 ±0.10	96.95 ±0.09	71.59 ±0.76
13	69.31 ±0.63	75.35 ±0.51	70.13 ±0.66	100	92.55 ±0.06	94.24 ±0.06	69.42 ±0.64
15	68.27 ±0.95	75.31 ±0.68	68.75 ±0.62	100	92.37 ±0.90	96.91 ±0.08	68.27 ±0.95
17	67.28 ±0.54	75.02 ±0.64	67.24 ±0.61	100	92.16 ±0.09	96.87 ±0.10	67.28 ±0.55
19	66.49 ±1.37	75.01 ±0.64	65.62 ±1.16	100	91.98 ±0.13	96.86 ±0.08	69.79 ±1.26
21	65.08 ±0.98	75.24 ±0.81	64.29 ±1.42	100	91.83 ±0.15	96.91 ±0.11	65.08 ±0.97

where, C1= Hyperthermophilic, C2= Thermophilic, C3= Psychrophilic, C4= Meosphilic, C1, C2, C3, & C4 are the mean accuracies for each class and corresponding standard deviation.

B. Classification of the Four Protein Classes Based on PACC Features

The results obtained for the K-NN classification for different K values based on the PACC features are presented in Table IV. The results shows that the proposed model has excellent accuracies especially for the "K" values of 1, 3, 5, respectively, then similar to the previous two models, the sensitivity started to suffer as the "K" values were increased. The highest achieved accuracy at K=1 was 98.75%, specificity 100%, and the sensitivity was 96% .

The data analysis showed that the performance of the proposed K-NN classifiers depends on the selected "K" value, as the "K" value increased, the performance of the four class model drops. This may be due to the size of the data which is so large in addition to the non-uniform distribution of the data. The results also revealed that PACC features were good indicators for the prediction of the thermal class of the four classes prokaryotic proteins. This is not surprising since some authors have successfully used PACC features before in protein classification to predict thermophiles and mesophiles [19], [36], and in the discrimination protein membranes [29].

TABLE IV. PERFORMANCE RESULTS OF THE PACC K-NN PREDICTIVE MODEL USING 70% TRAINING AND 30% TEST

K	C1 %	C2 %	C3 %	C4 %	Accuracy %	Specificity %	Sensitivity %
1	96.164 ±0.43	97.98 ±1.14	95.27 ±0.55	99.46 ±0.09	98.72 ±0.13	99.66 ±0.11	96.21 ±0.48
3	87.15 ±0.67	89.28 ±1.30	69.58 ±0.80	95.73 ±0.11	92.50 ±0.21	96.67 ±0.11	87.15 ±0.66
5	79.96 ±0.39	76.51 ±1.12	67.90 ±1.26	95.23 ±0.12	89.83 ±0.13	96.36 ±0.08	79.96 ±0.39
7	72.86 ±0.97	74.71 ±1.75	62.86 ±1.44	94.20 ±0.13	87.99 ±0.18	96.36 ±0.06	72.86 ±0.97
9	66.65 ±0.88	79.12 ±0.71	55.86 ±0.72	92.82 ±0.15	86.64 ±0.14	96.05 ±0.14	66.65 ±0.89
11	59.79 ±0.94	79.45 ±1.18	51.84 ±0.92	92.16 ±0.17	85.42 ±0.20	96.27 ±0.13	59.79 ±0.94
13	53.05 ±0.81	78.94 ±2.98	47.47 ±1.15	91.72 ±0.15	84.36 ±0.11	96.32 ±0.09	53.05 ±0.80
15	49.00 ±0.95	76.53 ±1.03	49.99 ±0.75	91.69 ±0.12	83.80 ±0.17	96.22 ±0.15	49.00 ±0.95
17	44.55 ±1.38	77.33 ±0.67	46.88 ±1.28	90.22 ±3.15	83.02 ±0.18	96.37 ±0.15	44.55 ±1.37
19	40.51 ±1.19	78.49 ±0.79	44.39 ±0.81	91.06 ±0.17	79.59 ±9.5	96.39 ±0.14	40.51 ±1.19
21	37.79 ±1.18	78.98 ±1.17	41.12 ±1.07	90.74 ±0.21	82.02 ±0.21	96.39 ±0.09	37.78 ±1.19

where, C1= Hyperthermophilic, C2= Thermophilic, C3= Psychrophilic, C4= Meosphilic, C1, C2, C3, & C4 are the mean accuracies for each class and corresponding standard deviation

IV. CONCLUSION

In this work, two predictive models based on K-NN classifier were proposed to discriminate between fours classes of prokaryotic proteins namely, hyperthermophiles, thermophiles, psychrophiles, and mesophiles using ACC and PAAC features, which were extracted from protein sequences. The models were able to efficiently identify the temperature group which a query protein belongs to without having to perform hideous and lengthy laboratory experiments. Future work will look at the discriminative power of other machine learning algorithms such as SVM, neural networks, and others to predict the four temperature classes. Moreover, future work will look at feature selection by combing several physio-chemical features that are most powerful in discriminating each class. In addition will attempt to select new subset features which are powerful in identifying each of the four groups.

ACKNOWLEDGMENT

Special thanks to the Public Authority of Applied Education and Training for supporting this project, and special thanks to Northumbria University for its collaboration and support. We would like to also thank the reviewing committee of the conference for taking the time to review this paper

REFERENCES

[1] Z. D. Stephens, S. Y. Lee, F. Faghri, R. H. Campbell, C. Zhai, M. J. Efron, *et al.*, "Big data: astronomical or genomical?" *PLoS Biol.* vol. 13, no. 7, p. e1002195, 2015.

[2] K. Chou, "Prediction of protein cellular attributes using pseudo - amino acid composition," *Proteins*, vol. 43, no. 3, pp. 246-255, 2001.

[3] K. Chou, "Some remarks on protein attribute prediction and pseudo amino acid composition," *Journal of Theoretical Biology*, vol. 273, no. 1, pp. 236-247, 2011.
[4] K. Chou, "Impacts of bioinformatics to medicinal chemistry," *Medicinal Chemistry*, vol. 11, no. 3 pp. 218-234, 2015.
[5] A. Razvi and J. M. Scholtz, "Lesson in stability from thermophilic proteins," *Protein Science*, vol. 15, pp. 1569-1578, 2016.
[6] S. Talluri, "Advances in engineering of proteins for thermal stability," *Int. J. of Adv. Biotech. and Research*, vol. 2, no. 1, pp. 190-200, 2011.
[7] 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, 2008.
[8] M. M. Gromiha, M. Oobatake, and A. Sarai, "Important amino acid properties for enhanced thermostability from mesophilic to thermophilic proteins," *Biophysical Chemistry*, vol. 82, no. 1, pp. 51-67, 1999.
[9] M. M. Gromiha and M. X. Suresh, "Discrimination of mesophilic and thermophilic proteins using machine learning algorithms," *Proteins*, vol. 70, no. 4, pp. 1274-1279, 2008.
[10] G. Zhang and B. Fang, "Application of amino acid distribution along the sequence for discriminating mesophilic and thermophilic proteins," *Process Biochemistry*, vol. 41, no. 8, pp. 1792-1798, 2006.
[11] G. Zhang and B. Fang, "Discrimination of thermophilic and mesophilic proteins via pattern recognition methods," *Process Biochemistry*, vol. 41, no. 3, pp. 552-556, 2006.
[12] S. Kumar and N. Nussinov, "How do the thermophilic proteins deal with heat," *Cellular and Molecular Life Sciences*, vol. 58, no. 9, pp. 1216-1233, 2001.
[13] S. Trivedi, "Protein thermostability in archaea and eubacteria," *Genetics and Molecular Research*, vol. 5, no. 4, pp. 816-827, 2006.
[14] G. Zhang and B. Fang, "Support vector machine for discrimination of thermophilic and mesophilic proteins based on amino acid composition," *Protein Pept. Lett.*, vol. 13, no. 10, pp. 965-970, 2006.
[15] H. Lin and W. Chen, "Prediction of thermophilic proteins using feature selection technique," *Journal of Microbiological Methods*, vol. 84, no. 1, pp. 67-70, 2011.
[16] S. Nakariyakul, Z. P. Liu, and L. Chen, "Detecting thermophilic proteins through selecting amino acid and dipeptide composition features," *Amino Acids*, vol. 42, no. 5, pp. 1947-1953, 2011.
[17] Y. C. Zuo, W. Chen, G. L. Fan, and Q. Z. Li, "A similarity distance of diversity measure for discriminating mesophilic and thermophilic proteins," *Amino Acids*, vol. 44, no. 2, pp. 573-580, 2013.
[18] L. Wang and C. Li, "Optimal subset selection of primary sequence features using the genetic algorithm for thermophilic proteins identification," *Biotechnol Let.*, vol. 36, no. 10, pp. 1963-1969, 2014.
[19] G. L. Fan, Y. L. Liu, and H. Wang, "Identification of thermophilic proteins by incorporating evolutionary and acid dissociation information into Chou's general pseudo amino acid composition," *Journal of Theoretical Biology*, vol, 407, pp. 138-142, 2016.
[20] L. C. Wu, J. X. Lee, H. D. Huang, B. J. Liu, and J. T. Horng, "An expert system to predict protein thermostability using decision tree," *Expert Syst. Appl.*, vol. 36, no. 5, pp. 9007-9014, 2009.
[21] A. Albayrak and U. Sezerman, "Discrimination of thermophilic and mesophilic proteins using reduced amino acid alphabets with n-grams," *Current Bioinformatics*, vol. 7, no. 2, pp. 152-158, 2012.
[22] 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, 2001.
[23] L. Montanucci, P. Fariselli, P. L. Martelli, and R. Casadio, "Predicting protein thermostability changes from sequence upon multiple mutations," *Bioinformatics*, vol. 24, no. 13, pp. i190-i195, 2008.
[24] 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, 2004.
[25] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, *et al.*, "The protein data bank," *Nucleic Acids Research*, vol. 28, no. 1, pp. 235-242, 2000.
[26] M. Mahmoudi, A. A. Arab, J. Zahiri, and Y. Parandian, "An overview of the protein thermostability prediction: Databases and tools," *Journal of Nanomedicine Research*, vol. 3, no. 6, p. 00072, 2016.
[27] H. P. Modarres, , M. R. Mofrad, and A. S. Nezhad, "ProtDataTherm: A database for thermostability analysis and engineering of proteins," *PLoS ONE*, vol. 13, no. 1, p. e0191222, 2018.
[28] Z. R. Li, H. H. Lin, L. Y. Han, L. Jiang, X. Chen, and Y. Z. Chen, "PROFEAT: A web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence," *Nucleic Acids Res.*, vol. 34, pp. w32-w37, 2016.
[29] D. Wang, L. Yang, Z. Fu, and J. Xia, "Prediction of thermophilic protein with pseudo amino acid composition: An approach from combined feature selection and reduction," *Protein Pept. Lett.*, vol. 18, no. 7, pp. 684-689, 2011.
[30] T. R. Golub, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science Journal*, vol. 286, no. 5439, pp. 531-537, 1999.
[31] S. S. Nanuwa and H. Seker, "Investigation into the role of sequence-driven-features for prediction of protein structural classes," in *Proc. of the 8th IEEE Int. Conf. in Bioinformatics and Bioengineering (BIBE2008)*, Athens, 2008, pp. 1-6.
[32] S. Kim, "Protein β-turn prediction using nearest-neighbors method," *Bioinformatics*, vol. 20, no. 1, pp. 40-44, 2004.
[33] S. Salzberg and S. Cost, "Prediction protein secondary structure with a nearest neighbor algorithm," *J. Mol. Biol.*, vol. 227, no. 2, pp. 371-374, 1992.
[34] A. A. Salamov and V. V. Solovyev, "Protein secondary structure prediction secondary structure by combining nearest-neighbor and multiple sequence alignments," *J. Mol. Biol.*, vol. 247, no. 1, pp. 11-15, 1995.
[35] F. Nigsch, A. Bender, B. V. Buuren, J. Tissen, E. Nigsch, and J. P. O. Mitchell, "Melting point prediction employing k-Nearest neighbors algorithms and genetics parameter optimization," *J. Chem. Inf. Model, American Chemical Society*, vol. 46, no. 6, pp. 412-2422, 2006.
[36] M. Hayat, and A. Khan, "Prediction of membrane protein types by using dipeptide and pseudo amino acid composition-based composite features," *IET Communications*, vol. 6, no. 18, pp. 3257-3264, 2012.

Hanaa M. Hussain obtained her B.S. degree in Electrical Engineering/Biomedical and Clinical Engineering from California State University, Long Beach, USA in Dec 2000. In Nov 2004, she obtained her MSc in Bioengineering from Strathclyde University in Glasgow, U.K. In Nov 2012, she obtained her PhD in Electronics Engineering from the University of Edinburgh, U.K. Her PhD thesis was on FPGA implementation of data mining and machine learning methods applied to Bioinformatics. Hanaa is currently interested in research related to healthcare technology, and data mining of Big Bio-data.

Hanaa worked as an electrical engineer at Kuwait National Petroleum Company (KNPC) between 2001-2003. After obtaining her MSc. in 2004, she worked in lecturing at the College of Technological Studies, Electronics Engineering Technology Department at The Public Authority of Applied Education and Training (PAAET) in Kuwait until 2007. Finally, in Nov 2012, she was appointed by the same employer in Kuwait as an ASSISTANT PROFESSOR in Electronics Engineering after obtaining her PhD from Edinburgh University. During the period from Sep 2018 till Sep 2018 she was on sabbatical visiting the department of Engineering and Environment at the University of Northumbria.

Dr. Hussain is a member in IEEE since 1996 including membership is IEEE Engineering in medicine and biology, and a member in the society of Kuwaiti Engineers. She also was a founder member of Kuwaiti Biomedical Engineering committee.

Huseyin Seker obtained his BSc degree in Electrical and Electronics Engineering from Selcuk University, Konya, Turkey. He then obtained his MSc from the Division of Biomedical Engineering, Department of Electronics and Communication Engineering, Istanbul Technical University, Turkey. Then he obtained his PhD from Coventry University, U.K.

Dr. Seker was the head of Bio-Health Informatics Research Group, Centre for Computational Intelligence, De Montfort University, Leicester,U.K., Dr. Seker was also a SENIOR LECTURER in the Informatics Department, Faculty of Technology at De Montfort University. Currently, Dr. Seker is the DIRECTOR OF ENTERPRISE & ENGAGEMENT AND READER in Computer Science Newcastle upon Tyne, Northumberland, United Kingdom. His research interests are in computational intelligence in bio and health informatics and computational biology.

Dr. Seker is a recipient of an international paper award from IEEE Engineering in Medicine and Biology Society.