Low Complexity Classification of Thermophilic Protein using One Hot Encoding as Protein Representation

Meredita Susanty[1], Rukman Hertadi[2], Ayu Purwarianti[3], Tati Latifah Erawati Rajab[4]*
School of Electrical Engineering and Informatics, Institut Teknologi Bandung, Bandung, Indonesia[1, 3, 4]
School of Computer Science, Universitas Pertamina, Jakarta, Indonesia[1]
Faculty of Mathematics and Natural Sciences, Institut Teknologi Bandung, Bandung, Indonesia[2]
Center for Artificial Intelligence (U-CoE AI-VLB), Institut Teknologi Bandung, Bandung, Indonesia[3]

***Abstract*—The laborious, and cost-inefficient biochemical methods for identifying thermophilic proteins necessarily require a rapid and accurate method for identifying thermophilic proteins. Recently, machine learning has become a more effective method for identifying specific classes of extremophiles. There is still a need for a low-cost method for identifying thermophilic proteins, despite the fact that studies employing machine learning yielded superior results to conventional methods. Here, we avoid the problem of manually crafted features, which involves experts defining and extracting a set of features using only protein sequences as input for various computational methods. This study classifies thermophilic proteins and their counterparts using only protein sequences in one-hot encoding representation and the bidirectional long short-term memory (BiLSTM) model. The model achieved an accuracy of 92.34 percent, a specificity of 91 percent, and a sensitivity of 93.77 percent, which is superior to other models reported elsewhere that rely on a number of manually crafted features. In addition, the more trustworthy and objective data set and the independent data set for evaluation make this model competitive with other, more accurate models.**

***Keywords*—Thermophilic; classification; one-hot encoding; BiLSTM**

I. Introduction

Extremophiles are microorganisms that have adapted to inhabit ecological niches deemed "extreme" due to unfavorable environmental conditions, such as excessively high or low temperatures, extreme pH values, high salt concentrations, or high pressure. Proteins isolated from extremophiles are biomolecules that possess unusual properties. One exceptional property enables proteins to function under extreme conditions. Researchers use extremophilic proteins in industrial applications since they can resist harsh conditions, which presents new opportunities for biocatalysts and biotransformation [1]–[4]. Thermophiles, extremophiles that thrive at temperatures between 41 and 122 degrees Celsius, with optimal growth temperatures between 60 and 108 degrees Celsius, are among the most studied.

Identifying thermophilic proteins' biochemical and physicochemical properties is critical because this serves as the foundation for designing and engineering proteins and enzymes. However, the biochemical identification method for thermophilic proteins is time-consuming, labor-intensive, and costly. Therefore, a rapid and accurate method for identifying thermophilic proteins is urgently required.

The increasing availability of data on extremophilic proteins enables computational methods to predict protein classification and identify the key characteristics that define that class [5]. These computational methods provide a more effective means of identifying specific classes of extremophiles than biochemical methods that require wet lab experiments. Many researchers have recently attempted to distinguish thermophilic organisms from their counterparts using machine learning [6]. In general, classifying thermophilic proteins using an approach based on machine learning involves six steps: dataset collection, data pre-processing, feature extraction, feature or dimension reduction, classification, and evaluation. Numerous techniques have been developed for researching and identifying thermophilic proteins. Kumar et al. (2000) identified and categorized thermophilic proteins based on structural differences at room temperature [7], whereas Gromiha et al. (2001) studied the properties of amino acids and the effects of amino acid residues on protein heat resistance [8]. These studies rely on expensive and ineffective biological techniques from the past. In other studies, Zhou et al. (2008) used an amino acid coupling model to identify thermophilic proteins [9]. Zhang et al. (2006) utilized dipeptide composition and amino acid composition to differentiate between mesophilic and thermophilic proteins [10] and achieved an accuracy of 86.6% for five-fold cross-validation. Using neural network-based amino acid composition [11], Gromiha and Suresh (2008) increased the computational complexity of five-fold cross-validation by 89%. Using decision tree methods, Wu et al. classified and identified thermophilic proteins with an accuracy of over 80% [12]. The accuracy of the k-nearest neighbor classifier used by Zuo et al. (2013) to classify thermophilic proteins was 91.02 percent [13].

To classify thermophile proteins and their counterparts, most previous studies have used hand-picked features calculated as input features for various machine learning models. During the feature extraction stage, various features are extracted using various computational methods based on amino acid sequence and transformed into numerical vectors. Protein features have been used in previous studies [6], [11]–

*Corresponding Author.

[19]. Even the most recent study [20], [21] still employs this type of protein representation. However, their scalability is limited because some of these input features are not always available and are computationally expensive to generate. Wu et al. (2009) predicted protein thermostability using protein structure and sequence characteristics, and suggested that although sequence and structural models had slightly higher accuracy, sequence-only models can provide sufficient accuracy for thermostability prediction [12]. Inspired by Wu's claim, we use only protein sequence as input to avoid the problem of obtaining handcrafted features, which requires experts to define and extract a set of features using various computational methods. This study improves accuracy by using sequence-only, removing all complexity to calculate derived features, and bidirectional long short-term memory (BiLSTM).

II. METHODS

The core structure of the present research consists of the five processes listed below: (1) dataset collecting, (2) feature extraction, (3) classification, and (4) performance evaluation. The framework's flowchart is presented in Fig. 1. One-hot representation is used to encode protein sequence. This representation is input to three distinct classifiers: Multi-layer Perceptron (MLP), Convolutional Neural Network (CNN), and Bidirectional Long Short-Term Memory (BiLSTM).

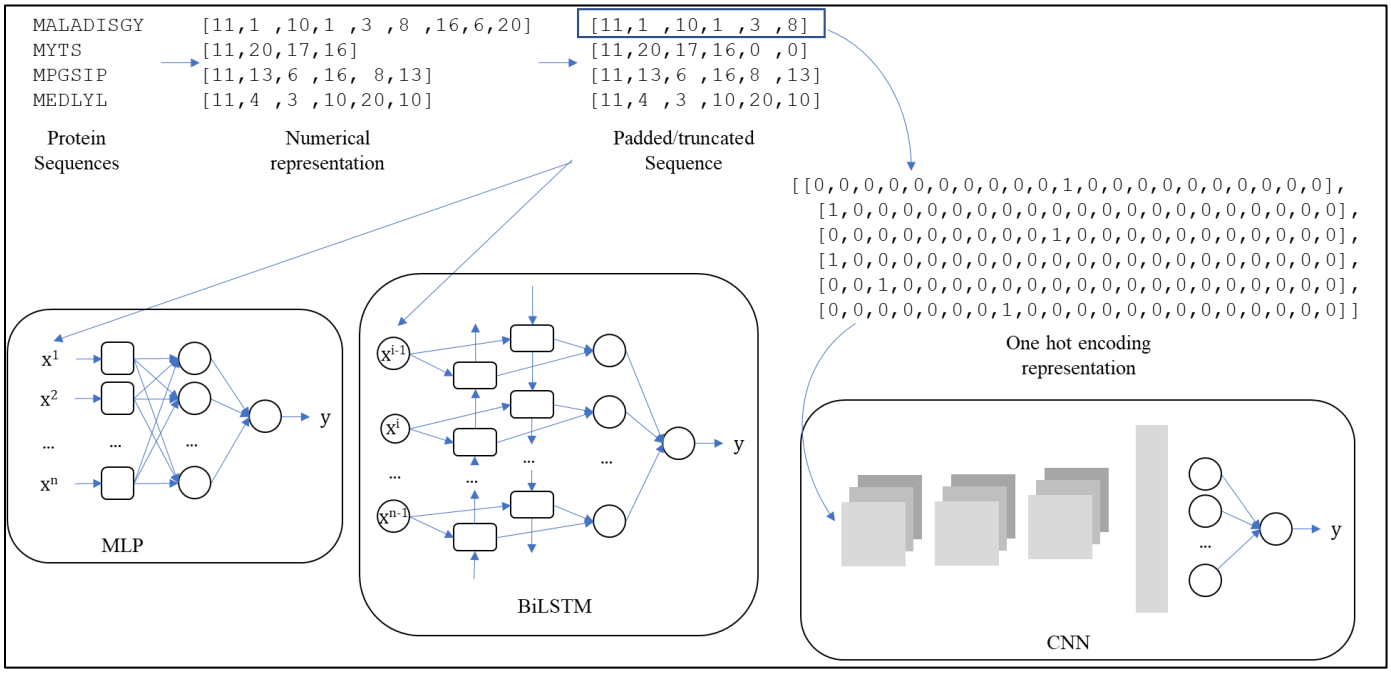

Fig. 1. Flow chart of a framework for predicting thermophilic proteins

A. Datasets

We utilized Ahmed et al. (2022) thermophilic and mesophilic benchmark dataset [21]. The dataset originated from the Universal Protein Resource (http://www.uniprot.org). To provide reliable data, previous studies retained only proteins that had been manually reviewed and excluded proteins with ambiguous residues, sequences that were fragments of other proteins, and proteins inferred from prediction or homology. With the CD-HIT algorithm [22] and a sequence identity threshold of 30%, redundancy and homology bias have been eliminated. There were 1,443 non-thermophilic proteins and 1,367 thermophilic proteins in the final benchmark dataset. The training and testing datasets are identical to those used in the study by Ahmed et al [21].

B. Feature Extraction

Here, instead of describing a protein with physical attributes, we directly encode its amino acid sequence. This approach of vectorizing categorical data is called one-hot encoding. An $L \times n$ matrix encodes a protein sequence of length L, where n is the number of amino acids. Each row of the matrix consists of $(n-1)$ 0s and a single 1, with the location of the 1 representing the amino acid residue in the protein at that position.

Using the constructed code dictionary, a 1 letter code is substituted for an integer value for each unaligned amino acid sequence. If the code is not included in the dictionary, the value is replaced by 0 and only the 20 most common amino acids are considered. This phase will transform the 1 letter code sequence data into numerical data. Next, post padding is performed with a maximum sequence length of 1024, either padding with 0 if the entire sequence length is less than 1024 or truncating the sequence to a maximum length of 1024.

C. Classification

We examined several classifiers, including MLP, CNN, and BiLSTM, to find the best model for classifying thermophilic proteins. **MLP** is a feed-forward neural network having input, hidden, and output layers, which are responsible for receiving, processing, and final prediction, respectively. The network is trained using backpropagation with the supervised learning technique. Each trained neuron's output is defined by equation (1), where x_i represents the firing neuron's input values, w_i their weights, f the activation function, and b the neuron's

activation threshold [23], [24]. In the current study, the hidden layer activation function was a rectified linear activation unit (ReLU), while the outer layer activation function was sigmoid. To train the model, input, hidden, and output layers with one neuron, respectively, were used. Table I. shows the hyperparameters in detail.

$$f(\propto) = f(\sum_{i=1} w_i x_i + b) \quad (1)$$

CNN employs layers with convolving filters applied to local features [25]. The use of convolutional neural networks is the same as that of picture data. The sole difference is that 1D convolutions are applied as opposed to 2D convolutions. The convolution network is made of the following: (1) A Conv1D with 128 units, with the relu activation function and a kernel size of two to extract basic properties, (2) another Conv1D with 128 units with relu activation function, kernel size of two, and l2 regularizer, (3) a dropout layer that randomly drops nodes during training, (4) a MaxPooling1D layer that down-samples the input by taking max over the steps that are constrained to a pool_size in each stride to reduce the spatial size of the representation, (5) another dropout layer, (6) a Dense layer with 250 units for the fully connected layer, and (7) an output layer with the sigmoid activation function because this is a binary problem. Table I. shows the hyperparameters in detail.

LSTMs are excellent at maintaining long-term memory dependencies. The LSTM architecture accomplishes this through the use of input gate, the forget gate, and the output gate [26]. The inputs of unidirectional LSTM, as we can see, are from the past. Hence, it only retains past knowledge. On the other hand, a bidirectional LSTM is able to preserve contextual information from both the past and the future at any time because it run the inputs in two directions, one from the past to the future and one from the future to the past [27]. The first layer of the model is the embedding layer which uses the 256-length vector, and the next layer is the bidirectional LSTM layer which has 256 neurons. These will work as the memory unit of the model, which has a vocab size of 21 representing 20 unique amino acids and one padding character. L2 regularization is added to prevent model over-fitting. After bidirectional LSTM, the dense layer is an output layer with sigmoid function. The hyperparameter used in this study is summarized in Table I.

D. Classification

In order to evaluate the overall model performance, the following parameters were used

$$Sn = \frac{TP}{TP+FN} \quad (2)$$

$$Sp = \frac{TN}{TN+FP} \quad (3)$$

$$Acc = \frac{TP+TN}{TP+FN+FP+TN} \quad (4)$$

$$F1 = 2 \times \frac{Sn \times (\frac{TP}{TP+FP})}{Sn+(\frac{TP}{TP+FP})} \quad (5)$$

where Sn, Sp, Acc, F1 denote sensitivity (or recall), specificity, accuracy, and F1-score. In this study we use the F1-score to make sure that when the accuracy is high, it correctly predicts both classes. Thermophilic proteins classified as thermophilic are designated as TP (true positive), non-thermophilic proteins labeled as non-thermophilic are TN (true negative), non-thermophilic proteins classed as thermophilic are FP (false positive), and thermophilic proteins regarded as non-thermophilic are FN (false negative).

III. Results and Discussion

In this study we solely used protein sequences to classify thermophilic proteins. This chain of amino acids was represented using one-hot encoding which is the input for the classifier. Three machine learning techniques, MLP, CNN, and BiLSTM were used to identify thermophilic and non-thermophilic proteins. Hyperparameters tuning was performed to find the optimal model. The best optimized hyperparameters for each model are described in Table I. The result of each model is shown in Table II. The result of each classifier is measured using the following performance metric; accuracy, specificity, recall, and F1-score. The macro average values of specificity, recall, and F1-score are presented in Table II.

TABLE I. Best Hyperparameter

Hyperparameter	MLP	CNN	BiLSTM
Number of Layer	2	4	3
Number of Neuron	100-1	250-1	256-256-1
Number of Filter	-	128	-
Activation Function	Relu	Relu	Relu
Optimizer	Adam	Adam	Adam
Learning Rate	0.0001	0.001	0.001
Dropout	-	0.6 - 0.8	-
Regularizer	0.4	0.9	0.1
Early Stopping	True	True	True
Batch Size	60	256	256
Epoch	250	250	500

TABLE II. Performance Result

Model	Sn (%)	Sp (%)	Acc (%)	F1(%)
MLP	2.19	97.23	51.06	4.18
CNN	90.10	89.96	90.03	89.76
BiLSTM	**93.77**	**91.00**	**92.34**	**92.25**

[a.] Note: Sn: sensitivity (or recall), Sp: specificity, Acc: accuracy, and F1: F1-score

A. Performance Comparison on Different Algorithms

We compared the performance of three machine learning methods: MLP, CNN, and BiLSTM. The same features were used to train and test these methods. Based on the result as shown in Table II, the highest accuracy, sensitivity, specificity, and f-measure were achieved using BiLSTM, at 92.34%, 93.77%, 91.00%, and 92.25% respectively. The MLP, on the other hand, was only capable of achieving an accuracy of 51.06%.

TABLE III. PERFORMANCE OF CNN ON VARIOUS K-MERS

k-mers	Sn (%)	Sp (%)	Acc (%)	F1(%)
1	61.53	53.63	57.47	58.43
2	82.78	89.27	86.12	85.28
3	**90.10**	**89.96**	**90.03**	**89.76**
4	79.85	94.11	87.18	85.82
5	86.81	89.27	88.07	87.61

[b.] Note: Sn: sensitivity (or recall), Sp: specificity, Acc: accuracy, and F1: F1-score

MLP can theoretically approximate any function to any precision. Using the same dataset, MLP architecture in other studies which used protein sequence and several other features as the input could achieve accuracy of 99.26% [21]. However, MLPs were not ideal for processing patterns with sequence. Because we rely on the sequence of amino acids in this study, MLP strives to remember patterns in sequential data to discover reliance on historical data, which is extremely useful for prediction.

CNN learns to recognize spatial patterns. CNN worked remarkably well, however, when applied to specific NLP issues [28]. When applying CNN to sequential data, the result of each convolution will be triggered when a unique pattern is identified. By altering the size of the kernels and concatenating their outputs, it is possible to identify patterns of numerous sizes, such as 2, 3, or 5 adjacent amino acids (k-mers), which are analogous to the term n-grams. Therefore, CNN can distinguish this k-adjacent amino acid regardless of its position in the sequence. The results of our experiments with various k-mers are summarized in Table III, which reveals that prediction using three consecutive residues (amino acid 3-mers) provides the highest performance with an accuracy of 90.03 percent, a sensitivity of 90.10 percent, a specificity of 89.96 percent, and an f-measure of 89.76 percent.

TABLE IV. PERFORMANCE COMPARISON

Studies	Features[c]	Algorithms[d]	Sn (%)	Sp (%)	Acc (%)
Liu[19]	AAC	SVM	98.88	1.0	99.44
Feng [15]	AAC, reduced DC, physicochemical	SVM	98.2	98.2	98.2
Ahmed[21]	AAC, tPseAAC, aPseAAC, CKSAAP, DC, DDE, CTD	MLP	96.34	96.16	96.26
Guo[29]	AAC, DC, DDE, CTDC, CTDT, CTriad, CKSAAP, GTPC, GDPC, TPC	SVM	96.22	95.85	96.02
Wang[17]	pseAAC, AAC, PC, CTD	SVM	96.17	95.69	95.93
Sunny and Saleena[30]	AA frequency, pI, protein binding domain, disorder regions, conserved residues, buried exposed regions	RF	95.53	96.01	95.71
Tang [31]	5-mers AA	SVM	94.8	94.1	94.4
Charoenkwan [20]	AAC, DPC, CTDC, CTDD, CTDT, AAI, aPseAAC, pseAAC, PSSM_COM, RPM_PSSM, S_FPSSM	Meta-predictor optimization	95.1	93.3	94.2
Fan et al.[14]	AAC,pseAAC,pKA, PSSM-400	SVM	89.50	95.64	93.53
Nakariyakul [16]	AAC, DC	SVM	93.0	93.7	93.3
Wang [32]	AAC, PCP, GDC, entropy density, autocorrelation coefficient	SVM	91.68	93.44	92.56
Zhang and Fang [33]	AA Sequence	SVM	92.8	92	92.4
BiLSTM (Our Model)	**AA sequence**	**BiLSTM**	**93.77**	**91.00**	**92.34**
Albayrak [34]	RAAA, N-grams	SVM	92.1	91.4	91.796
Zuo[13]	AAC-based similarity distance	KNN	88.37	92.24	91.02
Lin and Chen [35]	AAC, GDC	SVM	85.40	93.60	90.8
Gromiha and Suresh [11]	AAC	NN	83.30	92.00	89
Zhang and Fang [18]	AAC, DC	LogitBoost	87.34	90.77	88.94
Wu[12]	secondary and tertiary structural features	DT	73.1	88.5	83.6

[c.] amino acid composition (AAC), traditional pseudo amino acid composition (tPseAAC), amphiphilic pseudo amino acid composition (aPseAAC), pseudo amino acid composition (PseAAC), the composition of k-spaced amino acid pairs (CKSAAP), dipeptide composition (DC), dipeptide deviation from the expected mean (DDE), composition, transition, and distribution (CTD), CTD Composition (CTDC), CTD Transition (CTDT), Cojoint Triad (Ctriad), Grouped Dipeptide Composition (GTPC), Grouped Tripeptide Composition (GDPC), Tripeptide Composition (TPC) , Physicochemical Properties (PCP), g-gap Dipeptide Composition (GDC), Position Specific Scoring Matrices (PSSM), Physic Chemical (PC), Isoelectric Point (pI), Amino Acids (AA), Acid Dissociation Constant (pKa),Reduced Amino Acid Alphabet (RAAA), composition- transition-distribution -based features containing CTD (CTDC), distribution in CTD (CTDD), transition in CTD (CTDT), physi- cochemical property-based features containing amino acid index (AAI), evolutionary information-based features containing position-specific scoring matrix composition (PSSM_COM), PSSM of log-odds score of each amino acid in each posi- tion (RPM_PSSM) and PSSM based on the matrix transformation (S_FPSSM)

[d.] Support Vector Machine (SVM), Multilayer Perceptron (MLP), Random Forest (RF), Bidirectional Long Short-Term Memory (BiLSTM), K-Nearest Neighbour (KNN), Neural Network (NN), Decision Tree (DT)

Convolutions and pooling processes eliminate information about the local order of amino acids, making it more difficult to implement sequence tagging inside a pure CNN architecture. Additionally, pooling minimizes output dimensionality while maintaining the most essential information. Each filter recognizes a unique characteristic. If this specific feature appears somewhere in the sequence, applying the filter to that region will result in a large value, while applying the filter to other regions will result in a small value. By performing the max operation, we retain information regarding whether or not the feature existed in the sequence, but we lose information regarding where the feature appeared. Although it operates wonderfully, it is incapable of interpreting temporal data. In natural language processing, the resemblance between proteins and sentences explains why CNN performs worse than BiLSTM, which can learn from both past and future data in the sequence.

Among all models, the BiLSTM Model produces the best results. Because it knows what amino acids follow and precede an amino acid in the protein sequence, BiLSTM effectively increases the amount of information available to the network, improving the context available to the algorithm. BiLSTM mines the relationships between contexts in both directions, resulting in a better recognition for classifying thermophilic proteins.

B. Comparison to Other Models

Comparison with previously published methods is required to establish the superiority of the new method. It is widely known that the prediction results would be exaggerated if the suggested technique was trained and evaluated using a benchmark dataset containing a large proportion of homologous sequences. If two protein sequences possess a sequence similarity of greater than 40 percent, they are considered homologous. When a model is trained and evaluated on such a dataset, the vast majority of predictors always attain high levels of accuracy. For instance, previous studies [33] did not eliminate very identical or homologous sequences with 40% sequence identity from their datasets. Other studies only used a 40% cutoff eliminate the homologous sequence [11]. This study's dataset, which uses a 30% cutoff, is therefore judged to be more objective and dependable.

Numerous models have been developed to identify thermophilic proteins [6], [11], [14]–[17], [19], [29], [31]. All of the proposed models were developed using machine learning techniques and assessed using cross-validation. Nevertheless, our model was evaluated using independent data.

While state-of-the-art (SOTA) methods typically use various features derived from amino acid sequences as input to identify thermophilicity, the method(s) presented here use encoded single protein sequences as one-hot encoding, which serve as the only input feature for the prediction. One-hot encodings are sparse, memory inefficient, and high-dimensional by definition. In one-hot encoding, there is no concept of similarity between sequence or structure pieces; they are either identical or dissimilar. In Table IV we summarize the comparison with other models. From the comparison, we can see that using BiLSTM we can still achieve competitive results compared to other models which use various additional features as input, such as physicochemical properties, amino acid composition, and dipeptide compositions. This model even achieved better accuracy compared to another model from Wu et al.[12] which also consider protein structure in classifying the protein.

Although this study only uses the amino acid sequence without any derived features and combines it with the bidirectional sequential model, we can obtain a competitive classification result. Some other studies that use various handcrafted features provide higher performance. So, including the semantics information in the protein representation might increase the performance. In natural language processing, we can extract features using the Language Model (LM), referred to as embeddings. Instead of manually calculating each feature, LMs offer a potential alternative to this increasingly time-consuming database search as they extract features directly from single protein sequences [36]. Using embedding that has semantic information as input for classification can improve classification performance [37]. Use of embedding to represent proteins for classification as a downstream task can be an exciting topic to explore in the future.

IV. CONCLUSIONS

This research demonstrates that models trained using simply amino acid sequences perform comparably to and frequently exceed models trained using multiple feature representations. Comparing different machine learning algorithms demonstrates that a sequential model with a bidirectional mechanism is applicable to all protein attributes, and that the position of amino acids in the protein sequence has a significant predictive function. This research facilitates the development of predictive models by bypassing many of the challenges associated with generating the biological, chemical, and physical attributes that describe protein sequences.

ACKNOWLEDGMENT

This work was supported by Indonesian Ministry of Education, Culture, Research and Technology for Fiscal Year 2022.

REFERENCES

[1] K. Dumorné, D. C. Córdova, M. Astorga-Eló, and P. Renganathan, ‘Extremozymes: A Potential Source for Industrial Applications’, J. Microbiol. Biotechnol., vol. 27, no. 4, pp. 649–659, Apr. 2017, doi: 10.4014/JMB.1611.11006.

[2] M. De Champdoré, M. Staiano, M. Rossi, and S. D’Auria, ‘Proteins from extremophiles as stable tools for advanced biotechnological applications of high social interest’, J. R. Soc. Interface, vol. 4, no. 13, p. 183, Apr. 2007, doi: 10.1098/RSIF.2006.0174.

[3] S. Tapadar et al., ‘Role of Extremophiles and Extremophilic Proteins in Industrial Waste Treatment’, Remov. Emerg. Contam. Through Microb. Process., pp. 217–235, 2021, doi: 10.1007/978-981-15-5901-3_11.

[4] D. Zhu, W. A. Adebisi, F. Ahmad, S. Sethupathy, B. Danso, and J. Sun, ‘Recent Development of Extremophilic Bacteria and Their Application in Biorefinery’, Front. Bioeng. Biotechnol., vol. 8, p. 483, Jun. 2020, doi: 10.3389/FBIOE.2020.00483/BIBTEX.

[5] P. Charoenkwan, N. Schaduangrat, M. M. Hasan, M. A. Moni, P. Lió, and W. Shoombuatong, ‘Empirical Comparison and Analysis of Machine Learning-Based Predictors for Predicting and Analyzing of Thermophilic Proteins’, EXCLI J., vol. 21, pp. 554–570, 2022, doi: 10.17179/excli2022-4723.

[6] H. Lin and W. Chen, 'Prediction of thermophilic proteins using feature selection technique', J. Microbiol. Methods, vol. 84, no. 1, pp. 67–70, Jan. 2011, doi: 10.1016/J.MIMET.2010.10.013.

[7] S. Kumar, C. J. Tsai, and R. Nussinov, 'Factors enhancing protein thermostability', Protein Eng., vol. 13, no. 3, pp. 179–191, 2000, doi: 10.1093/PROTEIN/13.3.179.

[8] M. M. Gromiha, 'Important inter-residue contacts for enhancing the thermal stability of thermophilic proteins', Biophys. Chem., vol. 91, no. 1, pp. 71–77, Jun. 2001, doi: 10.1016/S0301-4622(01)00154-5.

[9] 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, doi: 10.1007/S00726-007-0589-X.

[10] G. Zhang and B. Fang, 'Application of amino acid distribution along the sequence for discriminating mesophilic and thermophilic proteins', Process Biochem., vol. 41, no. 8, pp. 1792–1798, Aug. 2006, doi: 10.1016/J.PROCBIO.2006.03.026.

[11] M. M. Gromiha and M. X. Suresh, 'Discrimination of mesophilic and thermophilic proteins using machine learning algorithms', Proteins Struct. Funct. Bioinforma., vol. 70, no. 4, pp. 1274–1279, Mar. 2008, doi: 10.1002/PROT.21616.

[12] L. C. Wu, J. X. Lee, H. Da 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, Jul. 2009, doi: 10.1016/J.ESWA.2008.12.020.

[13] 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, doi: 10.1007/s00726-012-1374-z.

[14] 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', J. Theor. Biol., vol. 407, pp. 138–142, 2016, doi: 10.1016/j.jtbi.2016.07.010.

[15] C. Feng, Z. Ma, D. Yang, X. Li, J. Zhang, and Y. Li, 'A Method for Prediction of Thermophilic Protein Based on Reduced Amino Acids and Mixed Features', Front. Bioeng. Biotechnol., vol. 8, no. May, pp. 1–10, 2020, doi: 10.3389/fbioe.2020.00285.

[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, 2012, doi: 10.1007/s00726-011-0923-1.

[17] 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, doi: 10.2174/092986611795446085.

[18] G. Zhang and B. Fang, 'LogitBoost classifier for discriminating thermophilic and mesophilic proteins', J. Biotechnol., vol. 127, no. 3, pp. 417–424, Jan. 2007, doi: 10.1016/J.JBIOTEC.2006.07.020.

[19] X.-L. Liu, J.-L. Lu, and X.-H. Hu, 'Predicting thermophilic proteins with pseudo amino acid composition:approached from chaos game representation and principal component analysis', Protein Pept. Lett., vol. 18, no. 12, pp. 1244–1250, Oct. 2011, doi: 10.2174/092986611797642661.

[20] P. Charoenkwan, N. Schaduangrat, M. A. Moni, P. Lio', B. Manavalan, and W. Shoombuatong, 'SAPPHIRE: A stacking-based ensemble learning framework for accurate prediction of thermophilic proteins', Comput. Biol. Med., vol. 146, no. June, p. 105704, 2022, doi: 10.1016/j.compbiomed.2022.105704.

[21] Z. Ahmed et al., 'iThermo: A Sequence-Based Model for Identifying Thermophilic Proteins Using a Multi-Feature Fusion Strategy', Front. Microbiol., vol. 13, p. 82, Feb. 2022, doi: 10.3389/FMICB.2022.790063/BIBTEX.

[22] Y. Huang, B. Niu, Y. Gao, L. Fu, and W. Li, 'CD-HIT Suite: a web server for clustering and comparing biological sequences', Bioinformatics, vol. 26, no. 5, pp. 680–682, Jan. 2010, doi: 10.1093/BIOINFORMATICS/BTQ003.

[23] M. Popescu, V. Balas, L. Perescu-Popescu, and N. Mastorakis, 'Multilayer perceptron and neural networks', WSEAS Trans. Circuits Syst. , vol. 8, no. 7, 2009, doi: 10.5555/1639537.1639542.

[24] M. Popescu, V. Balas, O. Olaru, N. Mastorakis, and O. Olaru, 'The Backpropagation Algorithm Functions for the Multilayer Perceptron', 2009.

[25] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, 'Gradient-based learning applied to document recognition', Proc. IEEE, vol. 86, no. 11, pp. 2278–2323, 1998, doi: 10.1109/5.726791.

[26] S. Hochreiter and J. Schmidhuber, 'Long Short-Term Memory', Neural Comput., vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.

[27] A. Graves and J. Schmidhuber, 'Framewise phoneme classification with bidirectional LSTM networks', Proc. Int. Jt. Conf. Neural Networks, vol. 4, pp. 2047–2052, 2005, doi: 10.1109/IJCNN.2005.1556215.

[28] Y. Kim, 'Convolutional Neural Networks for Sentence Classification', EMNLP 2014 - 2014 Conf. Empir. Methods Nat. Lang. Process. Proc. Conf., pp. 1746–1751, Aug. 2014, doi: 10.48550/arxiv.1408.5882.

[29] Z. Guo, P. Wang, Z. Liu, and Y. Zhao, 'Discrimination of Thermophilic Proteins and Non-thermophilic Proteins Using Feature Dimension Reduction', Front. Bioeng. Biotechnol., vol. 8, no. October, pp. 1–10, 2020, doi: 10.3389/fbioe.2020.584807.

[30] J. S. Sunny and L. M. Saleena, 'Amino acid frequency and domain features serve well for random forest based classification of thermophilic and mesophilic protein; a case study on serine proteases', 2021.

[31] 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. 4, pp. 1–8, 2017, doi: 10.1142/S1793524517500504.

[32] X.-F. Wang, P. Gao, Y.-F. Liu, H.-F. Li, and F. Lu, 'Predicting Thermophilic Proteins by Machine Learning', Curr. Bioinform., vol. 15, no. 5, pp. 493–502, 2020, doi: 10.2174/1574893615666200207094357.

[33] 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, Nov. 2006, doi: 10.2174/092986606778777560.

[34] A. Albayrak and U. O. Sezerman, 'Discrimination of Thermophilic and Mesophilic Proteins Using Reduced Amino Acid Alphabets with n-Grams', Curr. Bioinform., vol. 7, no. 2, pp. 152–158, 2012, doi: 10.2174/157489312800604435.

[35] H. Lin and W. Chen, 'Prediction of thermophilic proteins using feature selection technique', J. Microbiol. Methods, vol. 84, no. 1, pp. 67–70, 2011, doi: 10.1016/j.mimet.2010.10.013.

[36] D. Ofer, N. Brandes, and M. Linial, 'The language of proteins: NLP, machine learning & protein sequences', Comput. Struct. Biotechnol. J., vol. 19, pp. 1750–1758, 2021, doi: 10.1016/j.csbj.2021.03.022.

[37] A. Elnaggar et al., 'ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Deep Learning and High Performance Computing', bioRxiv, Jul. 2020, Accessed: Mar. 29, 2021. [Online]. Available: http://arxiv.org/abs/2007.06225.