Research Article GU J Sci, Part A, 11(4): 701-710 (2024) 10.54287/gujsa.1558391

Gazi University
Journal of Science
PART A: ENGINEERING AND INNOVATION
http://dergipark.org.tr/gujsa

Machine Learning Approaches for Differentiating Thermophilic and Mesophilic Lipases

Nurcan VARDAR-YEL[1*]

[1] *Altınbaş University, Vocational School of Health Services, Medical Laboratory Techniques Program, İstanbul, Türkiye*

Keywords

Mesophilic
Thermophilic
Lipase
Machine Learning

Abstract

Differentiating thermophilic proteins from their mesophilic counterparts presents a significant challenge, yet achieving this distinction is crucial for the rational design of more stable proteins. In this study, a systematic analysis was performed on 3,715 unreviewed bacterial lipase enzymes obtained from the UniProt web server and screened according to their Tm values. Furthermore, a tree was constructed using the MEGA 11 program and lipase sequences from different families were selected. The final dataset consists of 88 mesophilic proteins and 123 thermophilic proteins were used. We found that Ile, Leu, aliphatic index, hydropathy, aliphatic amino acids, hydrophobic amino acids, tiny amino acids, and small amino acids are the key variables distinguishing thermophilic from mesophilic lipase proteins. These findings suggest that amino acid composition is crucial in differentiating these two groups.

Cite

Vardar-Yel, N. (2024). Machine Learning Approaches for Differentiating Thermophilic and Mesophilic Lipases. *GU J Sci, Part A, 11*(4), 701-710. doi:10.54287/gujsa.1558391

Author ID (ORCID Number)

0000-0003-0994-5871 Nurcan VARDAR-YEL

Article Process

Submission Date 30.09.2024
Revision Date 14.10.2024
Accepted Date 25.10.2024
Published Date 30.12.2024

1. INTRODUCTION

One of the features that make protein thermostability an important issue in both biochemical and biotechnological research is its ability to increase reaction rate and efficiency at high temperatures. (Rigoldi et al., 2018). Unfortunately, few proteins are stable at high temperatures, resulting in the need for accurate methods to predict whether a protein is globally thermodynamically stable from its primary sequence. The key elements that are correlated with high protein thermostability are salt bridges, dipeptide patterns, ion pairs, and amino acid content (Kumar et al., 2000; Gromiha et al., 2002; Lin & Chen, 2011; Ahmed et al., 2022). For example, thermophilic proteins often have higher levels of residues like Ile, Arg, Glu, Lys, and Pro, while Ser, Asn, Gln, Thr, and Met are lower compared to mesophilic proteins (Gromiha & Suresh, 2008; Feng et al., 2020). As a result, the identification of protein thermostability-driving forces in sequence features have been utilized to develop methods which predict thermophilic properties from amino acid coupling patterns and dipeptide compositions (Das & Gerstein, 2000; Liang et al., 2005; Zhang & Fang, 2006a; 2006b; 2007; Lin & Chen, 2011). Additionally, single point mutations also modulate thermostability, which emphasizes the need for careful sequence analysis (Capriotti et al., 2004; 2005; Gromiha, 2007; Montanucci et al., 2008; Tian et al., 2010; Li et al., 2012; Marabotti et al., 2021). Numerous structural characteristics, including disulfide bonds, hydrophobic interactions, pi-pi and cation-pi interactions and salt bridges are fundamental in defining thermostability and should be evaluated when designing proteins and enzymes (Loladze et al., 1999; Razvi & Scholtz 2006; Strickler et al., 2006; Vardar-Yel et al., 2024). Thermophilic organisms are useful for industrial applications because they produce high-functioning enzymes and can withstand temperatures between 41°C and 122°C (Das & Gerstein, 2000). Proteins that are particularly thermophilic were searched for to identify the contributions of sequence and structural features associated with thermostability (Mrozek & Małysiak-

*Corresponding Author, e-mail: nurcan.vardar@altinbas.edu.tr

Mrozek, 2011; Dao et al., 2017; Charoenkwan et al., 2021). Increased levels of non-polar amino acids in thermostable enzyme structures are believed to enhance the hydrophobicity of proteins, which attracts them to the catalytic pocket and increases their rigidity. Furthermore, thermostable enzymes exhibit a greater number of hydrophobic and disulfide linkages. These characteristics facilitate conformational folding and provide enzymes with a more rigid structure (Li et al., 2005; Hussian et al., 2023). Additionally, stronger electrostatic contacts in a protein's outer regions lead to more ion pair interactions when amino acids have a greater charge. In thermophilic proteins as opposed to their mesophilic counterparts, these electrostatic forces play a larger role in maintaining the stability of the folded form. This shows that in order to keep proteins stable at high temperatures, electrostatic interactions are essential (Dominy et al., 2004; Hussian et al., 2023). In their study, Zhou et al. (2008) examined the differences in amino acid composition between mesophilic and thermophilic proteins, highlighting key features in thermophilic proteins such as increased hydrophobicity, decreased uncharged polar residues, elevated charged and aromatic residues, specific amino acid coupling patterns, and distinct amino acid preferences (Zhou et al., 2008). Other studies have isolated specific patterns of cavity dipeptide specific to thermophilic protein sequences (Wijma et al., 2013). This level of comparison is highly helpful since it provides detailed information about the major impacts of packing, hydrophobic interactions, disulfide bridges, and aromatic interactions on protein thermostability (Christensen & Kepp, 2013). The ability to distinguish between mesophilic and thermophilic proteins with accuracy has been demonstrated by machine learning methods such as support vector machines, decision trees, neural networks, and logistic functions (Ding et al., 2004; 2010; Zhang & Fang, 2006a; 2006b; 2007; Gromiha & Suresh, 2008; Lin & Chen, 2011; Ai et al., 2012; Albayrak & Sezerman, 2012; Chakravorty et al., 2017; Feng et al., 2020). In this study, two distinct algorithms, Support Vector Machines (SVM) and Decision Trees, have been used to analyze the differences between thermophilic and mesophilic lipase enzymes from bacteria that are unreviewed from the Uniprot database.

2. MATERIAL AND METHOD

2.1. The Dataset

3715 unreviewed bacterial lipase enzymes from the Uniprot web server were screened for Tm values. Redundancies were removed using the CD-HIT (Cluster Database at High Identity with Tolerance) algorithm and erroneous sequences were eliminated. Furthermore, a tree was constructed using the MEGA 11 program and lipase sequences from different families were selected (Tamura et al., 2021). The final dataset consists of 88 mesophilic proteins and 123 thermophilic proteins were used. Lipase enzymes screened from different thermophilic and mesophilic bacterial sources are listed in Table 1. Enzymes from various thermophilic and mesophilic bacterial organisms, identified under names such as triacylglycerol lipase, monoacylglycerol lipase, carboxylesterase, esterase, alpha/beta hydrolase, lipase, and Lipase EstA, were screened (Table 1). A comprehensive analysis was conducted on 38 distinct variables, including amino acid composition, sequence length, aliphatic index, instability index, net charge, hydropathy, molecular weight (Da), and the number of various amino acid groups. These groups consisted of charged (DEKHR), aliphatic (ILV), aromatic (FHWY), polar (DERKQN), neutral (AGHPSTY), hydrophobic (CFILMVW), positively charged (KRH), negatively charged (DE), as well as tiny (ACDGST), small (EHILKMNPQV), and large (FRWY) amino acids. Data was collected from the COPid-Calculate Composition of Whole Protein and Tm Predictor website (Kumar et al., 2008; Ku et al., 2009).

2.2. Cross Validation

The correctness of results determines the success of systems developed for any objective. The most common approach is k-fold cross-validation (Alataş et al., 2023). Here, an original dataset is divided into k subsets of roughly equal size. The system trains itself in k-1 subsets and tests itself in the remaining one. The hypothesis' validity is indicated by the average of the error value throughout the course of these k experiments.

For this experiment, a value of k equated to 5 was used as the size of the dataset was of medium scale, and more than that would have required additional computational power. This method has been used to compensate for the inadequacies of the test-train split method. Finally, the dataset was split into a training and testing set in a 7:3 ratio. The accuracy derived from this split was measured and the results were compared with those developed by the cross-validation method.

Table 1. *Lipase enzymes screened from different thermophilic and mesophilic bacterial sources*

Thermophilic protein	**Source**	**Mesophilic protein**	**Source**
Triacylglycerol lipase	*Geobacillus sp.* GHH01	Triacylglycerol lipase	*Streptococcus downei*
Triacylglycerol lipase	*Geobacillus stearothermophilus*	Triacylglycerol lipase	*Cupriavidus necator*
Triacylglycerol lipase	*Geobacillus thermoleovorans*	Lipase	*Rhodococcus* sp.
Monoacylglycerol lipase	*Thermus thermophilus*	Lipase1	*Streptomyces ambofaciens*
Monoacylglycerol lipase	*Geobacillus thermopakistaniensis*	Triacylglycerol lipase	*Mycolicibacterium fortuitum*
Monoacylglycerol lipase	*Chloroflexi bacterium*	Lipase EstA	*Limnohabitans* sp.
Monoacylglycerol lipase	*Thermoflexales bacterium*	Triacylglycerol lipase	*Pseudomonas fluorescens*
Carboxylesterase	*Geobacillus stearothermophilus*	Esterase/lipase lipF	*Mycobacterium tuberculosis*
Carboxylesterase	*Geobacillus thermodenitrificans*	Monoacylglycerol lipase	*Bacillus* sp.
Carboxylic ester hydrolase	*Geobacillus thermodenitrificans*	Lipase2	*Staphylococcus aureus*
GDSL-family esterase	*Geobacillus thermodenitrificans*	Lipase	*Pseudomonas* sp.
Triacylglycerol lipase	*Geobacillus thermoleovorans (Bacillus thermoleovorans)*	Triacylglycerol lipase	*Escherichia coli*
Triacylglycerol lipase	*Geobacillus zalihae*	Triacylglycerol lipase	*Staphylococcus epidermidis*
Esterase-lipase	*Thermochaetoides thermophila*	Triacylglycerol lipase	*Staphylococcus* sp.
Monoacylglycerol lipase	*Thermoflexales bacterium*	Triacylglycerol lipase	*Bacillus anthracis*
Monoacylglycerol lipase	*Thermosipho africanus*		
Alpha/beta hydrolase	*Aquifex aeolicus*		

2.3. Machine Learning Algorithms

The algorithms used for the study were chosen in relation to the dataset used, looking for algorithms that work well with the limited amount of data available, can deal effectively with imbalanced class problems, are robust to outliers, and can be applied to a large number of data structures. Furthermore, the machine learning applications used in this study were implemented using scikit-learn, a free software library for the Python programming language. The anaconda suite (https://www.anaconda.com/download), which includes a chosen collection of Python packages, is the simplest method to obtain Python, the core packages, and Jupyter Notebook.

2.3.1. Random Forest (RF)

Numerous tree classifiers are used in combined machine learning methods like Random Forest. Each tree votes once for the popular class, and the final classification result is calculated by adding up all of the tree classifier ratings. The characteristics of RF are high classification accuracy, robust tolerance to noise and outliers, and resistance to overfitting (Liu et al., 2012).

2.3.2. Decision Tree (DT)

Due to their structured, reliable and user-friendly nature, Decision Trees have been widely applied to both classification and regression problems. Other reasons supporting its popularity are its interoperability with other systems and the existence of understandable concepts. DTs are constructed in a top-down manner: they start with the most general data and are progressively specialized. The methodology applied for their construction and the starting point for tree building are the main factors taken into account in the application of DTs (Kotsiantis, 2013)

2.3.3. Performance Evaluation Metrics

Table 2 displays the confusion matrix of the multiple categorization results.

***Table 2**. The confusion matrix of the outcomes. Cell1=TN: True Negative; Cell2=FP: False Positive; Cell3=FN: False Negative; Cell4=TP: True Positive*

Predicted Values	0: Mesophilic	TN Cell 1	FP Cell 2
	1: Thermophilic	FN Cell 3	TP Cell 4
		0: Mesophilic	1: Thermophilic
	Real Values		

As a result, Equations 1-4 were used to calculate a range of metrics that were utilized to assess the performance of the algorithm. These metrics include accuracy, precision (P), recall (R), the F-score, and the area under the ROC curve (AUC).

$$Accuracy = \frac{TN + TP}{TP + TN + FP + FN} \tag{1}$$

$$P = \frac{TP}{TP + FP} \tag{2}$$

$$R = \frac{TP}{TP + FN} \tag{3}$$

$$F\ score = \frac{2(R * P)}{R + P} \tag{4}$$

In this context, the terms TP, FP, TN, and FN represent specific classifications within the data analysis process. TP means true positives, where thermophiles are positively identified as thermophiles. FP or false positives are where mesophiles are classified as thermophiles. TN stands for true negatives, which are the number of mesophiles correctly identified as mesophiles, while FN stands for false negatives, which are cases where thermophiles are classified as mesophiles. The AUC was calculated from a plot of the relationship between the false positive rate on the x-axis and the true positive rate on the y-axis. This allows a visual and quantitative assessment of how well the algorithm discriminates between thermophiles and mesophiles according to the thresholds used for classification.

3. RESULTS AND DISCUSSION

This study compared the effectiveness of different machine learning techniques in discriminating between mesophilic and thermophilic proteins with amino acid composition as the main feature for classification. Several criteria were used to compare the performance of machine learning algorithms in this study: precision (%), accuracy (%), F1 score (%), recall (%), AUC values (%) and corresponding ROC curves (%). In this study, multiclass classification was used to predict the dependent variables of mesophilic and thermophilic bacterial lipase groups. These two dependent variables were treated as two different outputs. The confusion matrices obtained from both machine learning methods are shown in Figure 1. The first and fourth cells show the correctly predicted values, while the remaining cells show the incorrectly predicted values.

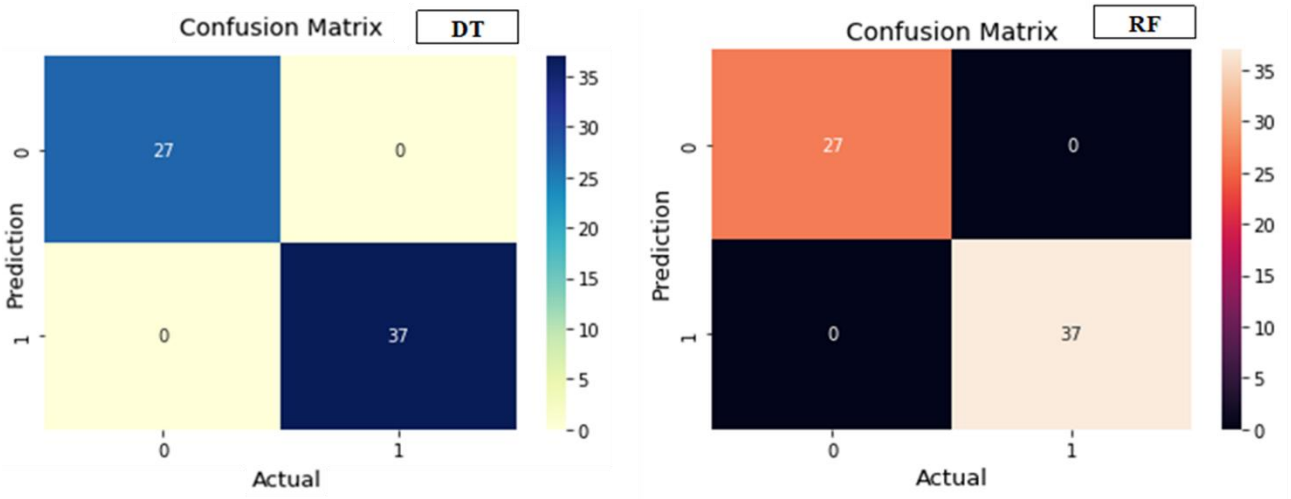

Figure 1. *Confusion matrix for RF and DT used in the prediction of mesophilic and thermophilic bacterial lipase enzyme.*

The results demonstrated that the majority of machine learning techniques exhibited an accuracy of approximately 99.5% in differentiating between thermophilic and mesophilic proteins. The results demonstrated that the random forest and decision tree approaches exhibited a high degree of similarity in this regard. Therefore, based on this finding, there is no significant difference in the performance of any of these machine learning algorithms.

ROC curves offer a very insightful approach for excellent visualization of the balance of sensitivity and specificity in a model with respect to the total predictive power of the classes presented in the model. For multi-class classification problems, an ROC curve can be constructed as a one-to-one approximation where each class is compared with all other classes. Thus, the number of ROC curves is equal to the number of classes used. In this case, the structure of the ROC curve was the same for both classes, indicating equal prediction performance. Figure 2 shows the ROC curve of DT.

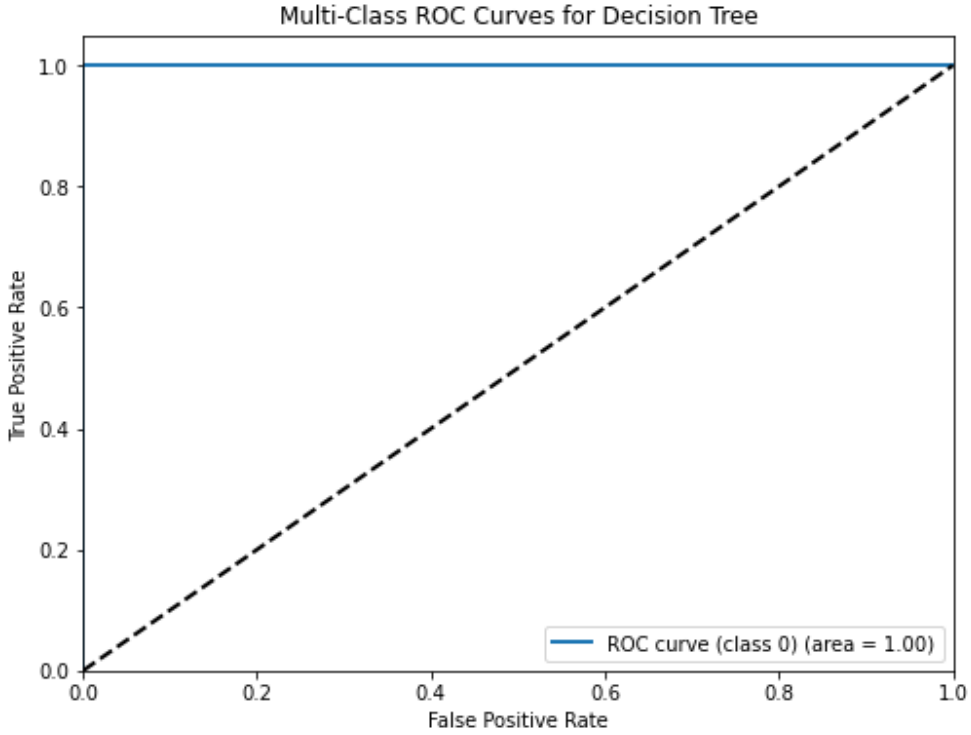

Figure 2. *ROC curve for classes 0, 1 and 2 with Decision Tree algorithms.*

A widely adopted approach for assessing model performance in classification tasks is the evaluation of accuracy, largely due to its simplicity and intuitive interpretation. Accuracy is calculated by taking the ratio of the total number of correct predictions to the overall number of predictions made by the model. Following the training-test split (7:3) and 10-fold cross-validation for the RF and DT algorithms, the accuracy values were calculated as 1. Following the evaluation of the ROC curve, AUC, and accuracy metrics, it was concluded that the Random Forest (RF) and Decision Tree (DT) algorithms demonstrated the highest effectiveness for machine learning classification of class 0 and class 1, corresponding to mesophilic and thermophilic categories within the dataset. As a result, the link between the features in the dataset and the living temperature classes were analyzed using the RF approach since both algorithms gave similar results. The feature importance values of the final model are shown in Figure 3.

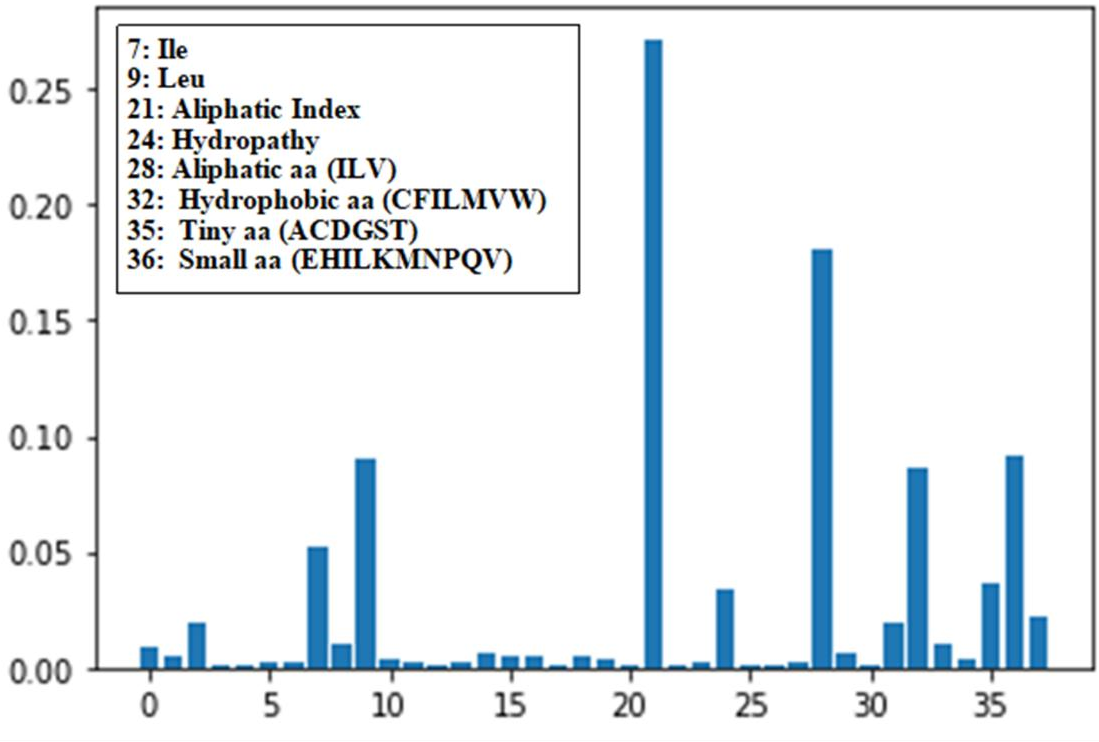

Figure 3. *Feature importance results for Random Forest*

The prediction involving thermophilic bacterial lipase enzymes resulted in the aliphatic index having the most dominant percentage contribution. This is probably to be expected because most of the analyzed protein data is normally filtered according to this parameter and thus proves that machine learning serves as a self-correcting function that produces excellent results. The aliphatic amino acids (I, L, V) also emerge as crucial, ranking second with a value of 18%. The association between the aliphatic index and aliphatic amino acid composition with thermostability is well-documented in numerous studies (Ikai, 1980; Ponnuswamy et al., 1982; Pack & Yoo, 2004; Wu et al., 2009; Sahoo et al., 2019).

Furthermore, the importance of hydrophobic amino acids (CFILMVW), Leu amino acid, and small amino acids (EHILKMNPQV) were found to be 8.7%, 9%, and 9%, respectively. Factors with lesser effects included Ile amino acid at 5.1%, hydropathy at 3.7%, and tiny amino acids (ACDGST) at 3.8%. These findings are consistent with those of previous studies. For example, Lin and Chen (2011) observed that individual amino acids such as glutamic acid, lysine and isoleucine play a crucial role in contributing to the thermostability of proteins. Indeed, related to this view, their research has shown that such amino acids significantly influence the structure of thermophilic proteins, allowing them to maintain their functionality at relatively high temperatures. In another study, Gromiha and Suresh (2008) compared the amino acid compositions between mesophilic and thermophilic proteins. They showed that compared to mesophilic proteins, thermophilic proteins had more charged residues such as Lys, Arg, Glu and Asp. They also found that among the hydrophobic residues, especially Val and Ile were more abundant in thermophilic proteins than in mesophilic proteins.

Numerous studies have emphasized that amino acid composition, dipeptide composition are distinguishing factor between thermophilic, mesophilic and psychrophilic proteins (Ding et al., 2004; 2010; Dominy et al., 2004; Zhang & Fang, 2006a; 2006b; 2007; Gromiha & Suresh, 2008; Lin & Chen, 2011; Ai et al., 2012; Albayrak & Sezerman, 2012; Chakravorty et al., 2017; Feng et al., 2020; Wang et al., 2020; Charoenkwan et al., 2021). However, a major difference in our study lies in the fact that we concentrated specifically on mesophylic and thermophilic bacterial-originated lipase proteins and non-protein types in general. In this way, our focused approach enabled an exploration that is going to be much more in-depth with regard to certain property and characteristic changes experienced by lipase proteins within these different thermal environments.

The results obtained in the lipase proteins and in the machine-learning studies need to strongly correlate with each other, which generally goes on to show that the mesophilic, thermophilic, and psychrophilic proteins present a high correlation among them. It can thus be stated that the data in question provides compelling evidence of the precision and reliability of the results obtained. The fact that our findings are consistent with some of the most advanced computational studies validates our methodology of research and further testifies to how amino acid compositions feature prominently in dictating thermal stability and functionality in lipase enzymes.

4. CONCLUSION

This study identifies the effective factors distinguishing mesophilic and thermophilic bacterial lipase enzymes. The results revealed that both machine learning algorithms demonstrated nearly identical accuracy, achieving a ten-fold cross-validation accuracy of 99% on a dataset consisting of 3,715 unreviewed bacterial lipase entries. According to feature importance results, Ile, Leu, aliphatic index, hydropathy, aliphatic amino acids, hydrophobic amino acids, tiny amino acids, and small amino acids are variables able to differentiate the thermophilic from the mesophilic lipase proteins. Therefore, this observation may indicate the role of amino acid composition in the differentiation of these two groups. The results obtained align with the results of previous studies comparing mesophilic and thermophilic proteins by machine learning.

ACKNOWLEDGEMENT

The author would like to thank Assoc. Prof. Dr. Handan Tanyıldızı Kökkülünk for her support in the analysis of machine learning data.

CONFLICT OF INTEREST

The author declares no conflict of interest.

REFERENCES

Ahmed, Z., Zulfiqar, H., Tang, L., & Lin, H. (2022). A statistical analysis of the sequence and structure of thermophilic and non-thermophilic proteins. *International Journal of Molecular Sciences*, *23*(17), 10116. https://doi.org/10.3390/ijms231710116

Alataş, E., Tanyıldızı Kökkülünk, H., Tanyıldızı, H., & Alcın, G. (2023). Treatment prediction with machine learning in prostate cancer patients. *Computer Methods in Biomechanics and Biomedical Engineering*, 1–9. https://doi.org/10.1080/10255842.2023.2298364

Albayrak, A., & Sezerman, U. O. (2012). Discrimination of thermophilic and mesophilic proteins using reduced amino acid alphabets with n-grams. *Current Bioinformatics*, *7*(2), 152-158. https://doi.org/10.2174/157489312800604435

Ai, H., Zhang, L., Zhang, J., Cui, T., Chang, A. K., & Liu, H. (2018). Discrimination of thermophilic and mesophilic proteins using support vector machine and decision tree. *Current Proteomics*, *15*(5), 374-383. https://doi.org/10.2174/1570164615666180718143606

Capriotti, E., Fariselli, P., & Casadio, R. (2004). A neural-network-based method for predicting protein stability changes upon single point mutations. *Bioinformatics*, *20*, 63-68. https://doi.org/10.1093/bioinformatics/bth928

Capriotti, E., Fariselli, P., & Casadio, R. (2005). I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Research*, *33*(2), 306-310. https://doi.org/10.1093/nar/gki375

Chakravorty, D., Faheem Khan, M., & Patra, S. (2017). Thermostability of proteins revisited through machine learning methodologies: From nucleotide sequence to structure. *Current Biotechnology*, *6*(1), 39-49. https://doi.org/10.2174/2211550105666151222183232

Charoenkwan, P., Chotpatiwetchkul, W., Lee, V. S., Nantasenamat, C., & Shoombuatong, W. (2021). A novel sequence-based predictor for identifying and characterizing thermophilic proteins using estimated propensity scores of dipeptides. *Scientific Reports*, *11*(1), 23782. https://doi.org/10.1038/s41598-021-03293-w

Christensen, N. J., & Kepp, K. P. (2013). Stability mechanisms of a thermophilic laccase probed by molecular dynamics. *PloS One*, *8*(4), e61985. https://doi.org/10.1371/journal.pone.0061985

Dao, F.-Y., Yang, H., Su, Z.-D., Yang, W., Wu, Y., Hui, D., Chen, W., Tang, H., & Lin, H. (2017). Recent advances in conotoxin classification by using machine learning methods. *Molecules*, *22*(7), 1057. https://doi.org/10.3390/molecules22071057

Das, R., & Gerstein, M. (2000). The stability of thermophilic proteins: a study based on comprehensive genome comparison. *Functional & Integrative Genomics*, *1*(1), 76-88. https://doi.org/10.1007/s101420000003

Ding, Y., Cai, Y., Zhang, G., & Xu, W. (2004). The influence of dipeptide composition on protein thermostability. *FEBS Letters*, *569*(1-3), 284-288. https://doi.org/10.1016/j.febslet.2004.06.009

Ding, Y. R., Cai, Y. J., Sun, J., & Xu, B. W. (2010). Identifying the mesophilic and thermophilic proteins from their amino acid composition with ν-support vector machines. *Journal of Algorithms & Computational Technology*, *4*(3), 335-348. https://doi.org/10.1260/1748-3018.4.3.335

Dominy, B. N., Minoux, H., & Brooks III, C. L. (2004). An electrostatic basis for the stability of thermophilic proteins. *Proteins: Structure, Function, and Bioinformatics*, *57*(1), 128-141. https://doi.org/10.1002/prot.20190

Feng, C., Ma, Z., Yang, D., Li, X., Zhang, J., & Li, Y. (2020). A method for prediction of thermophilic protein based on reduced amino acids and mixed features. *Frontiers in Bioengineering and Biotechnology*, *8*, 285. https://doi.org/10.3389/fbioe.2020.00285

Gromiha, M. M. (2007). Prediction of protein stability upon point mutations. *Biochemical Society Transactions*, *35*(6), 1569-1573. https://doi.org/10.1042/BST0351569

Gromiha, M. M., & Suresh, M. X. (2008). Discrimination of mesophilic and thermophilic proteins using machine learning algorithms. *Proteins: Structure, Function, and Bioinformatics*, *70*(4), 1274-1279. https://doi.org/10.1002/prot.21616

Gromiha, M. M., Thomas, S., & Santhosh, C. (2002). Role of cation-π interactions in the stability of thermophilic proteins. *Preparative Biochemistry and Biotechnology*, *32*(4), 355-362. https://doi.org/10.1081/PB-120015459

Hussian, C. H. A. C., & Leong, W. Y. (2023). Thermostable enzyme research advances: a bibliometric analysis. *Journal of Genetic Engineering and Biotechnology*, *21*(1), 37. https://doi.org/10.1186/s43141-023-00494-w

Ikai, A. (1980). Thermostability and aliphatic index of globular proteins. *The Journal of Biochemistry*, *88*(6), 1895-1898. https://doi.org/10.1093/oxfordjournals.jbchem.a133168

Kotsiantis, S. B. (2013). Decision trees: a recent overview. *Artificial Intelligence Review*, *39*(4), 261-283. https://doi.org/10.1007/s10462-011-9272-4

Ku, T., Lu, P., Chan, C., Wang, T., Lai, S., Lyu, P., & Hsiao, N. (2009). Predicting melting temperature directly from protein sequences. *Computational Biology and Chemistry*, *33*(6), 445-450. https://doi.org/10.1016/j.compbiolchem.2009.10.002 (The Tm Index program is available at http://tm.life.nthu.edu.tw/)

Kumar, S., Tsai, C.-J., & Nussinov, R. (2000). Factors enhancing protein thermostability. *Protein Engineering, Design and Selection*, *13*(3), 179-191. https://doi.org/10.1093/protein/13.3.179

Kumar, M., Thakur, V., & Raghava, G. P. S. (2008). COPid: composition based protein identification. *In Silico Biology*, *8*(2), 121-128. (Calculate Composition of Whole Protein is available at https://webs.iiitd.edu.in/raghava/COPid/whole_comp.html)

Liang, H.-K., Huang, C.-M., Ko, M.-T., & Hwang, J.-K. (2005). Amino acid coupling patterns in thermophilic proteins. *Proteins: Structure, Function, and Bioinformatics*, *59*(1), 58-63. https://doi.org/10.1002/prot.20386

Li, W. F., Zhou, X. X., & Lu, P. (2005). Structural features of thermozymes. *Biotechnology advances*, *23*(4), 271-281. https://doi.org/10.1016/j.biotechadv.2005.01.002

Li, Y., Zhang, J., Tai, D., Russell Middaugh, C., Zhang, Y., & Fang, J. (2012). PROTS: A fragment-based protein thermo-stability potential. *Proteins: Structure, Function, and Bioinformatics*, *80*(1), 81-92. https://doi.org/10.1002/prot.23163

Lin, H., & Chen, W. (2011). Prediction of thermophilic proteins using feature selection technique. *Journal of Microbiological Methods*, *84*(1), 67-70. https://doi.org/10.1016/j.mimet.2010.10.013

Liu, Y., Wang, Y., & Zhang, J. (2012, September 14-16). *New machine learning algorithm: Random Forest*. In: B. Liu, M. Ma, & J. Chang (Eds.), Proceedings of the Information Computing and Applications (pp. 246-252), Chengde, China. https://doi.org/10.1007/978-3-642-34062-8_32

Loladze, V. V., Ibarra-Molero, B., Sanchez-Ruiz, J. M., & Makhatadze, G. I. (1999). Engineering a thermostable protein via optimization of charge−charge interactions on the protein surface. *Biochemistry*, *38*(50), 16419-16423. https://doi.org/10.1021/bi992271w

Marabotti, A., Scafuri, B., & Facchiano, A. (2021). Predicting the stability of mutant proteins by computational approaches: An overview. *Briefings in Bioinformatics*, *22*(3), bbaa074. https://doi.org/10.1093/bib/bbaa074

Montanucci, L., Fariselli, P., Martelli, P. L., & Casadio, R. (2008). Predicting protein thermostability changes from sequence upon multiple mutations. *Bioinformatics*, *24*(13), i190-i195. https://doi.org/10.1093/bioinformatics/btn166

Mrozek, D., & Małysiak-Mrozek, B. (2011). An improved method for protein similarity searching by alignment of fuzzy energy signatures. *International Journal of Computational Intelligence Systems*, *4*(1), 75-88. https://doi.org/10.2991/ijcis.2011.4.1.7

Pack, S. P., & Yoo, Y. J. (2004). Protein thermostability: structure-based difference of amino acid between thermophilic and mesophilic proteins. *Journal of Biotechnology*, *111*(3), 269-277. https://doi.org/10.1016/j.jbiotec.2004.01.018

Ponnuswamy, P. K., Muthusamy, R., & Manavalan, P. (1982). Amino acid composition and thermal stability of proteins. *International Journal of Biological Macromolecules*, *4*(3), 186-190. https://doi.org/10.1016/0141-8130(82)90049-6

Razvi, A., & Scholtz, J. M. (2006). Lessons in stability from thermophilic proteins. *Protein Science*, *15*(7), 1569-1578. https://doi.org/10.1110/ps.062130306

Rigoldi, F., Donini, S., Redaelli, A., Parisini, E., & Gautieri, A. (2018). Engineering of thermostable enzymes for industrial applications. *APL Bioengineering*, *2*(1), 011501. https://doi.org/10.1063/1.4997367

Sahoo, R. K., Sanket, A. S., Gaur, M., Das, A., & Subudhi, E. (2019). Insight into the structural configuration of metagenomically derived lipase from diverse extreme environment. *Biocatalysis and Agricultural Biotechnology*, *22*, 101404. https://doi.org/10.1016/j.bcab.2019.101404

Strickler, S. S., Gribenko, A. V., Gribenko, A. V., Keiffer, T. R., Tomlinson, J., Reihle, T., Loladze, V. V., & Makhatadze, G. I. (2006). Protein stability and surface electrostatics: a charged relationship. *Biochemistry*, *45*(9), 2761-2766. https://doi.org/10.1021/bi0600143

Tamura, K., Stecher, G., & Kumar, S. (2021). MEGA11: molecular evolutionary genetics analysis version 11. *Molecular Biology and Evolution*, *38*(7), 3022-3027. https://doi.org/10.1093/molbev/msab120

Tian, J., Wu, N., Chu, X., & Fan, Y. (2010). Predicting changes in protein thermostability brought about by single- or multi-site mutations. *BMC Bioinformatics*, *11*, 1. https://doi.org/10.1186/1471-2105-11-370

Vardar-Yel, N., Tütüncü, H. E., & Sürmeli, Y. (2024). Lipases for targeted industrial applications, focusing on the development of biotechnologically significant aspects: A comprehensive review of recent trends in protein engineering. *International Journal of Biological Macromolecules*, *273*, 132853. https://doi.org/10.1016/j.ijbiomac.2024.132853

Wang, X.-F., Gao, P., Liu, Y.-F., Li, H.-F., & Lu, F. (2020). Predicting thermophilic proteins by machine learning. *Current Bioinformatics*, *15*(5), 493-502. https://doi.org/10.2174/1574893615666200207094357

Wijma, H. J., Floor, R. J., & Janssen, D. B. (2013). Structure-and sequence-analysis inspired engineering of proteins for enhanced thermostability. *Current Opinion in Structural Biology*, *23*(4), 588-594. https://doi.org/10.1016/j.sbi.2013.04.008

Wu, L.-C., Lee, J.-X., Huang, H.-D., Liu, B.-J., & Horng, J.-T. (2009). An expert system to predict protein thermostability using decision tree. *Expert Systems with Applications*, *36*(5), 9007-9014. https://doi.org/10.1016/j.eswa.2008.12.020

Zhang, G., & Fang, B. (2006a). Application of amino acid distribution along the sequence for discriminating mesophilic and thermophilic proteins. *Process biochemistry*, *41*(8), 1792-1798. https://doi.org/10.1016/j.procbio.2006.03.026

Zhang, G., & Fang, B. (2006b). Discrimination of thermophilic and mesophilic proteins via pattern recognition methods. *Process Biochemistry*, *41*(3), 552-556. https://doi.org/10.1016/j.procbio.2005.09.003

Zhang, G., & Fang, B. (2007). LogitBoost classifier for discriminating thermophilic and mesophilic proteins. *Journal of Biotechnology*, *127*(3), 417-424. https://doi.org/10.1016/j.jbiotec.2006.07.020

Zhou, X.-X., Wang, Y.-B., Pan, Y.-J., & Li, W.-F. (2008). Differences in amino acids composition and coupling patterns between mesophilic and thermophilic proteins. *Amino Acids*, *34*(1), 25-33. https://doi.org/10.1007/s00726-007-0589-x