EXCLI Journal 2022;21:554-570 – ISSN 1611-2156
Received: January 28, 2022, accepted: February 21, 2022, published: March 02, 2022

Review article:

EMPIRICAL COMPARISON AND ANALYSIS OF MACHINE LEARNING-BASED PREDICTORS FOR PREDICTING AND ANALYZING OF THERMOPHILIC PROTEINS

Phasit Charoenkwan[a], Nalini Schaduangrat[b], Md Mehedi Hasan[c], Mohammad Ali Moni[d], Pietro Lió[e], Watshara Shoombuatong[b,*]

[a] Modern Management and Information Technology, College of Arts, Media and Technology, Chiang Mai University, Chiang Mai, Thailand, 50200

[b] Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok, Thailand, 10700

[c] Tulane Center for Biomedical Informatics and Genomics, Division of Biomedical Informatics and Genomics, John W. Deming Department of Medicine, School of Medicine, Tulane University, New Orleans, LA 70112, USA

[d] School of Health and Rehabilitation Sciences, Faculty of Health and Behavioural Sciences, the University of Queensland, St Lucia, QLD 4072, Australia

[e] Department of Computer Science and Technology, University of Cambridge, Cambridge, CB3 0FD, UK

* **Corresponding author:** Watshara Shoombuatong, Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok, Thailand, 10700. Phone: +66 2 441 4371; Fax: +66 2 441 4380; E-mail: watshara.sho@mahidol.ac.th

https://dx.doi.org/10.17179/excli2022-4723

ABSTRACT

Thermophilic proteins (TPPs) are critical for basic research and in the food industry due to their ability to maintain a thermodynamically stable fold at extremely high temperatures. Thus, the expeditious identification of novel TPPs through computational models from protein sequences is very desirable. Over the last few decades, a number of computational methods, especially machine learning (ML)-based methods, for *in silico* prediction of TPPs have been developed. Therefore, it is desirable to revisit these methods and summarize their advantages and disadvantages in order to further develop new computational approaches to achieve more accurate and improved prediction of TPPs. With this goal in mind, we comprehensively investigate a large collection of fourteen state-of-the-art TPP predictors in terms of their dataset size, feature encoding schemes, feature selection strategies, ML algorithms, evaluation strategies and web server/software usability. To the best of our knowledge, this article represents the first comprehensive review on the development of ML-based methods for in *silico* prediction of TPPs. Among these TPP predictors, they can be classified into two groups according to the interpretability of ML algorithms employed (i.e., computational black-box methods and computational white-box methods). In order to perform the comparative analysis, we conducted a comparative study on several currently available TPP predictors based on two benchmark datasets. Finally, we provide future perspectives for the design and development of new computational models for TPP prediction. We hope that this comprehensive review will facilitate researchers in selecting an appropriate TPP predictor that is the most suitable one to deal with their purposes and provide useful perspectives for the development of more effective and accurate TPP predictors.

Keywords: Thermophilic protein, bioinformatics, classification, machine learning, feature representation, feature selection

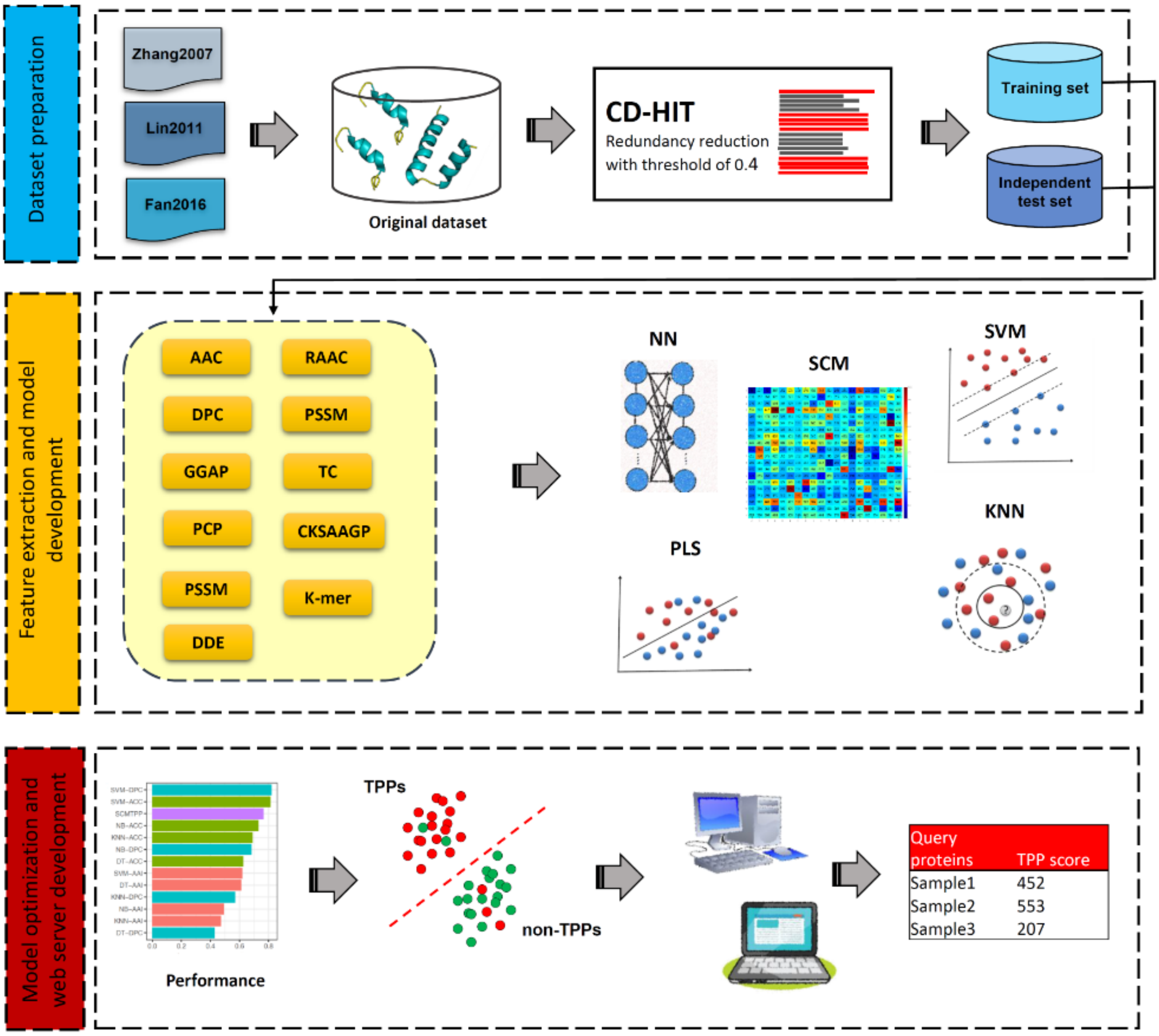
Received: January 28, 2022, accepted: February 21, 2022, published: March 02, 2022

INTRODUCTION

Proteins perform varied functions in the body such as enzyme catalysis, ion and molecular transport, antibody production, and cellular/physiological activity regulation and thus, are considered as one of the most important biological macromolecules. The three-dimensional structure of proteins heavily influences their functioning (Burley et al., 2017). Furthermore, structure-based drug design heavily relies on complex protein inter-residue interactions such as mechanisms of protein folding, rates of folding and unfolding, stability of protein structure, stability upon mutation, recognition mechanisms of protein-protein, protein-nucleic acid and protein-ligand complexes (Gromiha, 2010; Gromiha et al., 2019). Moreover, the critical role of Thermophilic proteins (TPPs) in biotechnology and chemical processing have already been established (Haki and Rakshit, 2003). TPPs maintain their stability at high temperatures (80-100 °C) as well as in the environmental temperatures of the host organism (Gaucher et al., 2008; Gromiha et al., 1999). Additionally, the stability of TPPs depends upon a variety of amino acid properties such as shape, hydration energy change (Gibbs function) in native proteins, dipeptide composition, amino acid residue contacts, ion pair numbers, hydrogen bonds, packing, and aromatic clusters (Gromiha et al., 1999; Pica and Graziano, 2016). Out of all the aforementioned properties, TPP stability relies mostly on hydrophobicity as the most important feature, followed by ion pairs and hydrogen bonds (Gromiha and Nagarajan, 2013). Therefore, in order to design proteins for specific medical or industrial applications, a thorough understanding of the molecular basis of protein thermostability is critical (Gromiha et al., 2019). Furthermore, the ease of TPP purification and their ability to withstand long periods of industrial conditions comes from their natural resistance to denaturation by chemical compounds (i.e., detergents, surfactants, oxidizing agents, and proteases) (Diaz et al., 2011; Habbeche et al., 2014; Huang et al., 2012b). Of note, survival of therapeutic proteins in blood is extended with higher thermostability (Narasimhan et al., 2010). Several advantages of TPPs include reduced contamination, mixes easily with low viscous agents while maintaining a high mass transfer rate as well as achieving greater solubility of products and substrates (Vieille and Zeikus, 2001). Furthermore, TPPs are advantageous in high-temperature pelleting processes (Rodriguez et al., 2000) as well as in the isomerization of glucose through endothermic reactions to generate high fructose syrups (Xu et al., 2014). Although experimental methods are the gold standard in verifying thermostability of proteins, these methods are usually labor-intensive, time-consuming and expensive. Thus, the rapid and accurate identification of TPPs from a large collection of proteins is highly advantageous and cost-effective.

Over the last few decades, a number of computational methods, especially machine learning (ML)-based methods, for *in silico* prediction of TPPs have been developed. The development of all these existing TPP predictors involves three main phases as summarized in Figure 1. The 1st phase is dataset preparation to form training and independent datasets. The 2nd phase is feature extraction and feature optimization. The 3rd phase is to train and evaluate a prediction model. The independent dataset is used to validate the effectiveness and robustness of the prediction model. Finally, the optimal prediction model is selected to establish a web server. We categorize the existing TPP predictors as listed in Table 1 into two groups according to the interpretability of the ML algorithms employed. The first group are the computational black-box methods, and there are nine out of fourteen existing TPP predictors (i.e., Gromiha et al.'s method (Gromiha and Suresh, 2008), ThermoPred (Lin and Chen, 2011), Wang et al.'s method (2011), Nakariyakul et al.'s method (2012), KNN-ID (Zuo et al., 2013), PSSM400_pKa (Fan et al., 2016), Tang et al.'s method (2017), Li et al.'s method (2019) and Feng et al.'s method (2020)) in this group.

Received: January 28, 2022, accepted: February 21, 2022, published: March 02, 2022



Query proteins	TPP score
Sample1	452
Sample2	553
Sample3	207

Figure 1: The overall framework of TPP predictors using machine learning methods. The 1st phase is dataset preparation to form training and independent datasets. The 2nd phase is feature extraction and feature optimization. The 3rd phase is to train and evaluate a prediction model. The independent dataset is used to validate the effectiveness and robustness of the prediction model. Finally, the optimal prediction model is selected to establish a web server.

The second group are the computational white-box methods, and there are five out of fourteen existing TPP predictors (i.e., Zhang et al.'s method (Zhang and Fang, 2006), LogitBoost (Zhang and Fang, 2007), Wu et al.'s method (2009), GA-MLR (Wang and Li, 2014) and SCMTPP (Charoenkwan et al., 2022)) in this group.

To the best of our knowledge, this article represents the first comprehensive review on the development of ML-based methods for *in silico* prediction of TPPs. In this study, our aim is to conduct an empirical comparison and analysis of fourteen existing TPP predictors in terms of multiple perspectives, including their feature encoding schemes, feature selection strategies, ML algorithms, evaluation strategies and web server/software usability as summarized in Table 1. First, we reviewed available training and independent datasets employed for developing the current TPPs predictors. The detailed information of these datasets are provided in Table 2. Second, the performance of various TPP predictors on two benchmark datasets (i.e., the Gromiha2007 (Gromiha and Suresh, 2008) and

Lin2011 (Lin and Chen, 2011)) and two independent datasets (i.e., the Zhang2007 (Zhang and Fang, 2006) and Charoenkwan2021 (Charoenkwan et al., 2022)) were compared and discussed. Our comparative results demonstrate that ThermoPred outperformed the competing TPP predictors in terms of both predictive performance and community utility, while SCMTPP outperformed the competing TPP predictors in terms of high interpretability and simplicity. Finally, we discuss the advantages and disadvantages of the current TPP predictors and provided future perspectives for the design and development of new computational models for TPP prediction.

MATERIALS AND METHODS

Framework of TPP prediction using machine learning-based approaches

The overall framework of TPP predictors using machine learning methods involves three main phases as summarized in Figure 1. The 1st phase is to prepare the high-quality dataset to generate training (for cross-validation and parameter optimization purposes) and independent (for assessing and validating the transferability and reliability) datasets. The 2nd phase is feature extraction and feature optimization. Feature extraction works to represent each protein sequence to capture the key information of TPPs and non-TPPs. The detailed information of feature encodings employed is recorded in Table 1. Since each protein sequence is represented as a high dimensional feature vector, it is well-known that the feature optimization step might help to exclude irrelevant/noisy features and lead to the improved performance of the trained model. Thus, the 3rd phase is to train and evaluate the prediction model. The independent dataset is used to validate the effectiveness and robustness of the prediction model. Finally, the optimal prediction model is selected to establish a web server. The details of web server availability and usability for TPP prediction is recorded in Table 1.

Datasets

Detailed information of all the training and independent datasets used for developing the existing methods are recorded in Table 2. Among these datasets, the Gromiha2007 (Gromiha and Suresh, 2008) (used for developing Gromiha et al.’s method (Gromiha and Suresh, 2008), KNN-ID (Zuo et al., 2013), and PSSM400_pKa (Fan et al., 2016)) and Lin2011 (Lin and Chen, 2011) (used for developing ThermoPred (Lin and Chen, 2011), Nakariyakul et al.’s method (Nakariyakul et al., 2012), GA-MLR (Wang and Li, 2014), Tang et al.’s method (Tang et al., 2017), Li et al.’s method (Li et al., 2019) and Feng et al.’s method (Feng et al., 2020)) datasets were two well-known training datasets used for developing almost all of the existing methods. As described in an article (Gromiha and Suresh, 2008), the training dataset of Gromiha2007 were directly derived from the Zhang2007 dataset (3521 TPPs and 4895 non-TPPs) and TPPs and non-TPPs with more than 40 % sequence identity were then excluded using the CD-HIT program. Finally, the training dataset of the Gromiha2007 dataset contained 1609 TPPs and 3075 non-TPPs. In case of the Lin2011 dataset (915 TPPs and 793 non-TPPs), Lin et al. collected TPPs and non-TPPs from 136 prokaryotic organisms extracted from the Universal Protein Resource (UniProt). Unfortunately, only the Lin2011 dataset can be accessed at http://lin-group.cn/server/ThermoPredv1. Recently, our group constructed an up-to-date dataset from several previous studies (Fan et al., 2016; Lin and Chen, 2011; Zhang and Fang, 2006) consisting of 6579 TPPs. After excluding redundant sequences using the CD-HIT program, 1823 TPPs and 3124 non-TPPs were obtained (called the Charoenkwan2021 dataset). The Charoenkwan2021 dataset can be downloaded at http://pmlabstack.pythonanywhere.com/SCMTPP.

State-of-the-art computational approaches for TPP prediction

More than ten ML-based approaches have been developed for TPP prediction. These approaches were developed using a variety of aspects, including the benchmark datasets, feature descriptors, feature section methods and ML algorithms, etc. In Table 1, we summarize 13 existing sequence-based TPP predictors along with their employed feature encoding schemes, ML algorithms and evaluation strategies. Most TPP predictors were trained and constructed in a five-step manner, which involves data preparation, feature extraction, feature selection, model optimization and development and web server construction. These existing sequence-based TPP predictors are categorized into two classes according to the interpretability of ML algorithms employed (Kurgan et al., 2009; Liang et al., 2021; Shoombuatong et al., 2017), which are described in detail below.

Table 1: A list of currently available machine learning-based methods for TPP identification summarized in this review

Method	Year	Classifier[a]	Features[b]	Evaluation strategy[c]	Web server availability status
Zhang et al.'s method (Zhang and Fang, 2006)	2006	PLS	AAC	Holdout	No
LogitBoost (Zhang and Fang, 2007)	2007	LogitBoost	AAC	5CV/IND	No
Gromiha et al.'s method (Gromiha and Suresh, 2008)	2008	NN	AAC	5CV/IND	No
Wu et al.'s method (Wu et al., 2009)	2009	DT	PCP	10CV	No
ThermoPred (Lin and Chen, 2011)	2011	SVM	AAC, GGAP	LOOCV	Yes, active
Wang et al.'s method (Wang et al., 2011)	2011	SVM	AAC, PCP, CTD	LOOCV	No
Nakariyakul et al.'s method (Nakariyakul et al., 2012)	2012	SVM	AAC, DPC	5CV-10CV/ LOOCV/IND	No
KNN-ID (Zuo et al., 2013)	2013	KNN	AAC	LOOCV/IND	Yes, inactive
GA-MLR (Wang and Li, 2014)	2014	MLR	AAC, GGAP	5CV/IND	No
PSSM400_pKa (Fan et al., 2016)	2016	SVM	AAC, pka, PSSM	10CV/IND	No
Tang et al.'s method (Tang et al., 2017)	2017	SVM	k-mer	5CV	No
Li et al.'s method (Li et al., 2019)	2019	Voting	AAC, DDE, TC, CKSAAGP	10CV	No
Feng et al.'s method (Feng et al., 2020)	2020	SVM	ACC, DPC, PCP,RAAC	10CV/IND	No
SCMTPP (Charoenkwan et al., 2022)	2021	SCM	GGAP	10CV/IND	Yes, active

[a] DT: decision tree, KNN: k-nearest neighbor, MLR: multiple linear regression, NN: neural networks, PLS: partial least-square regression, SCM: scoring card method, SVM: support vector machine

[b] AAC: amino acid composition, CKSAAGP: composition of kspaced amino acid group pairs, CTD: Composition-Transition-Distribution, DDE: Dipeptide deviation from expected mean, DPC: dipeptide composition, DPS: dipeptide propensity score; GGAP: g-gap dipeptide composition, k-mer: fragment-based technique; pka: acid dissociation constant, PCP: physicochemical properties, PseACC: pseudo amino acid composition, PSSM: position specific scoring matrix, RACC: reduce amino acid composition, TC: tripeptide composition

[c] 5CV: 5-fold cross-validation, 10CV: 10-fold cross-validation, IND: independent test, LOOCV: leave-one-out cross-validation

EXCLI Journal 2022;21:554-570 – ISSN 1611-2156
Received: January 28, 2022, accepted: February 21, 2022, published: March 02, 2022

Table 2: A summary of the training and independent test datasets used for developing the existing TPP predictors

Dataset[a]	CD-HIT threshold	Training dataset		Independent test dataset		Dataset availability
		Number of TPPs	Number of non-TPPs	Number of TPPs	Number of non-TPPs	
Zhang2006	1.0	76	81	20	20	Yes
Zhang2007	1.0	3521	4895	382	325	No
Gromiha2007	0.4	1609	3075	382	325	No
Wu2009	1.0	580	878	-	-	No
Lin2011	0.4	915	793	-	-	Yes
Wang2011	0.25	209	209	-	-	No
Nakariyakul2012 [b,d]	0.4	915	793	76	81	No
Zuo2013 [c,d]	0.4	1609	3075	76	81	No
Wang2014 [b,d]	0.4	915	793	76	81	No
Fan2016 [c,d]	0.4	1609	3075	76	81	No
Tang2017 [b]	0.4	915	793	-	-	No
Li2019 [b]	0.4	915	793	-	-	Yes
Feng2020 [b]	0.4	915	793	106	101	Yes
Charoenkwan2021	0.4	1482	1482	371	371	Yes

[a] Datasets' names are created by using the family name of the first author along with the publication year from the corresponding literature.
[b] Training dataset was directly obtained from the Lin2011 dataset
[c] Training dataset was directly obtained from the Gromiha2007 dataset
[d] Independent dataset was directly obtained from the Zhang2006 dataset

Performance evaluation and evaluation strategy

The predictive performance of our proposed model, baseline models and the two state-of-the-art methods is evaluated and compared using five common performance measures as follows: accuracy (ACC), sensitivity (Sn), specificity (Sp), Matthew's Correlation Coefficient (MCC) and area under the receiver-operating curves (AUC) (Azadpour et al., 2014; Charoenkwan et al., 2021d). These performance measures are described by the following equations:

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{(\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN})} \tag{1}$$

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

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

where TP, TN, FP and FN represent the number of true positives, true negatives, false positive and false negatives, respectively (Basith et al., 2020; Shoombuatong et al., 2017; Su et al., 2020a). It is well-known that Sn and Sp measure the predictive ability for two classes: positive and the negative, respectively. ACC, MCC and AUC evaluate the overall performance of the predictive model.

STATE-OF-THE-ART COMPUTATIONAL APPROACHES FOR TPP PREDICTION

More than ten ML-based approaches have been developed for TPP prediction. These approaches were developed using a variety of aspects, including the benchmark datasets, feature descriptors, feature section methods and ML algorithms, etc. In Table 1, we summarize 13 existing sequence-based TPP predictors along with their employed feature encoding schemes, ML algorithms and evaluation strategies. Most TPP predictors were trained and constructed in a five-step manner, which involves data preparation, feature extraction, feature selection, model optimization and development and web server construction. These existing sequence-based TPP predictors are categorized into two classes according to the interpretability of ML algorithms employed (Kurgan et al., 2009; Liang et al., 2021; Shoombuatong et al., 2017), which are described in detail below.

Prediction methods based on computational black-box methods

Among several ML algorithms used in this aspect, KNN (Zuo et al., 2013), NN (Gromiha and Suresh, 2008) and SVM (Fan et al., 2016; Feng et al., 2020; Lin and Chen, 2011; Nakariyakul et al., 2012; Tang et al., 2017; Wang et al., 2011) are known as Black-box approaches. Kurgan et al. (2009) described black-box methods as ML methods that cannot directly determine which features provide essential contribution to the prediction performance. As can be seen from Table 1, there are 6 out of 14 existing sequence-based predictors that were trained and developed by using SVM method (i.e., ThermoPred (Lin and Chen, 2011), Wang et al.‘s method (2011), Nakariyakul et al.’s method (2012), PSSM400_pKa (Fan et al., 2016), Tang et al.’s method (2017), Li et al.’s method (2019) and Feng et al.’s method (2020)). SVM has been successfully applied to solve variant research questions in computational biology and bioinformatics. The basic idea of SVM is to map the given input features into a high-dimensional space using kernel functions and find a maximum margin hyperplane that can separate positive samples from negative samples with a minimal misclassification rate (Chen et al., 2016; Manavalan and Lee, 2017). There are three well-known and commonly used kernel functions in SVM, including gaussian, polynomial and radial basis function (RBF). Particularly, the development of SVM models involves the optimization of two critical parameters, i.e. C and γ represent the regularization parameter and kernel parameter, respectively (Arif et al., 2020; Charoenkwan et al., 2020, 2021d).

In 2011, Lin et al. developed the first SVM-based approach for identifying TPPs (called ThermoPred) by using the Lin2011 dataset containing 915 TPPs and 793 non-TPPs. ThermoPred was trained with two feature descriptors (i.e., AAC and GGAP). In order to improve the predictive performance of ThermoPred, ANOVA technique was used to determine informative g-gap dipeptides. As a result, the informative g-gap dipeptides consisted of EE, KE, EI, I-K, I-E, E–K, E–E, K–E, Q–A and E---K, where – represents the gap of residues. In addition, the Sn, Sp and ACC of ThermoPred were 82.4 %, 93.0 % and 89.4 %, respectively, based on 5-fold cross-validation test. In the same year, Wang et al. developed another SVM-based approach for identifying TPPs. Their SVM-based approach was trained with three types of feature descriptors, including AAC, CTD and PCP. In addition, three feature selection methods (i.e., filter method, relief algorithm and genetic algorithm) were employed and used to determine informative features. Amongst these three feature selection methods, the highest ACC of 95.93 % was achieved by using genetic algorithm. The informative features derived from the genetic algorithm contained A, Q, I, K, F, Y, AA, AD, AQ, AS, RI, RK, DA, DQ, EE, EK, GQ, GI, GS, IN, IV, LY, MI, PA, SA, SQ, TI, YV, VY and CTD10. Recently, Feng et al. proposed another SVM-based approach trained with four feature descriptors (i.e., ACC, DPC, PCP and RAAC) (Feng et al., 2020). Then, principal component analysis was used to reduce irrelevant features and the final feature set contained 12 informative features.

In case of other ML algorithms employed, Gromiha et al. (Gromiha and Suresh, 2008) and Zuo et al. (2013) applied NN-based (called Gromiha et al.’s method) and KNN-based (called KNN-ID) models, respectively, trained with AAC to develop TPP predictors. For the 10-fold cross-validation results, ACC, Sn, and Sp of Gromiha et al.'s method (Gromiha and Suresh, 2008) were 89.00 %, 83.30 % and 92.00 %, respectively, while ACC, Sn, and Sp of KNN-ID (Zuo et al., 2013) were 90.66 %, 88.37 % and 92.24 %, respectively. To improve the accurate prediction of TPPs, Li et al. (2019) employed an ensemble strategy. To the best of the authors' knowledge, there is only one TPP predictor in existence that was constructed by using the ensemble strategy. Several previous studies have indicated that ensemble-based models are effective to provide improved performance over single-based models (Basith et al., 2022; Hasan et al., 2020; Kabir et al., 2022; Liang et al., 2021; Manavalan et al., 2019a; Rao et al., 2020). Their ensemble model provided a cross-validation ACC of 93.03 %.

Prediction methods based on computational white-box methods

Unlike black-box methods, white-box models are able to determine which features provide essential contribution to the prediction performance, such as DT (Wu et al., 2009), MLR (Wang and Li, 2014), PLS (Zhang and Fang, 2006) and SCM (Charoenkwan et al., 2022). Amongst several white-box methods, SCM has been indicated to achieve comparable performance to those of black-box methods, such as NN and SVM (Charoenkwan et al., 2013, 2021c, e, 2022). Huang et al. first introduced the original SCM method (Huang et al., 2012a), while Charoenkwan et al. developed an improved version by integrating both global and local sequence information (Charoenkwan et al., 2021c). The contribution of the SCM method is summarized in the following three aspects. First, the SCM method can discriminate positives from negatives by using only a single threshold value, emphasizing its ease-of-use and interpretability. Second, since the SCM method is known as a single feature-based model, indicating that this method could achieve better computational efficiency as compared to complex methods, such as SVM and ensemble approaches. Third, the SCM-derived propensity scores of 20 amino acids and 400 dipeptides are useful for characterizing and analyzing various functions of proteins and peptides.

In 2006, Zhang et al. proposed the first sequence-based predictor based on PLS algorithm for identifying TPPs based on 76 TPPs and 76 non-TPPs. Their method had the highest ACC for TPPs and non-TPPs prediction, which was 75 % and 85 %, respectively. Most recently, our group developed a new sequence-based predictor (called SCMTPP) for identifying and characterizing TPPs using estimated propensity scores of dipeptides. Furthermore, we established an up-to-date and high-quality dataset containing 1853 TPPs and 3233 non-TPPs from several published literatures. SCMTPP was developed using SCM method in conjunction with GGAP. SCMTPP based on the propensity scores of GGAP (g=0) was beneficial for TPP prediction with ACC of 88.30 %, MCC of 0.766 and AUC of 0.926 as evaluated by 10-fold cross-validation test. When compared with popular ML methods (i.e., DT, KNN and naive Bayes (NB)) on the training dataset, it could be noticed that SCMTPP outperformed those of DT-based, KNN-based and NB-based classifiers. Remarkably, SCMTPP's ACC was >7.05 %, >3.78 % and >1.86 % higher than DT-based, KNN-based and NB-based models, respectively.

RESULTS AND DISCUSSION

Comparative results on 5-fold and 10-fold cross-validation tests using the Gromiha2007 and Lin2011 datasets

In this section, we evaluated and compared the performance of different TPP predictors in terms of ACC, Sn and Sp using the two benchmark datasets (i.e., Gromiha2007 and Lin2011). To be specific, the Gromiha2007 dataset was used to evaluate five out of the fourteen existing TPP predictors (i.e., Gromiha et al.'s method, Wu et al.'s method, ThermoPred, KNN-ID, and PSSM400_pKa), while the Lin2011 dataset was used to evalute the six out of the fourteen existing TPP predictors (i.e., Gromiha et al.'s method, ThermoPred, Nakariyakul et al.'s method, GA-MLR, Tang et al.'s method and Feng et al.'s method).

The performance evaluation results of these two benchmark datasets are summarized in Figures 2 and Tables 3-4. As seen in Figure 2 and Table 3, ThermoPred outperformed the four competing TPP predictors (i.e., Gromiha et al.'s method, Wu et al.'s method, KNN-ID, and PSSM400_pKa) in terms of model complexity based on the Gromiha2007 dataset. Taking into consideration the cross-validation performance, PSSM400_pKa achieved the best performance in terms of ACC (93.53 %), Sn (89.50 %) and Sp (95.64 %) (Figure 2A). In the meanwhile, ThermoPred achieved the second-best performance in terms of ACC (93.53 %) and Sp (95.64 %). In terms of model complexity, ThermoPred (10D) performed better than PSSM400_pKa (460D) (Figure 2B). One of the major limitations of PSSM400_pKa was that there was no web server provided for this study. Therefore, its utility is limited to experimental scientists. In case of the Lin2011 dataset, Table 4 shows that ThermoPred still outperformed competing TPP predictors (i.e., Gromiha et al.'s method, Nakariyakul et al.'s method, GA-MLR, Tang et al.'s method and Feng et al.'s method) in terms of model complexity (Figure 2C-2D). It could be noticed that Feng et al.'s method achieved the best performance in terms of ACC (98.20 %), Sn (98.20 %) and Sp (98.20 %), while GA-MLR achieved the second-best performance in terms of the three performance metrics (i.e., ACC (95.61 %), Sn (95.41 %) and Sp (95.84 %)).

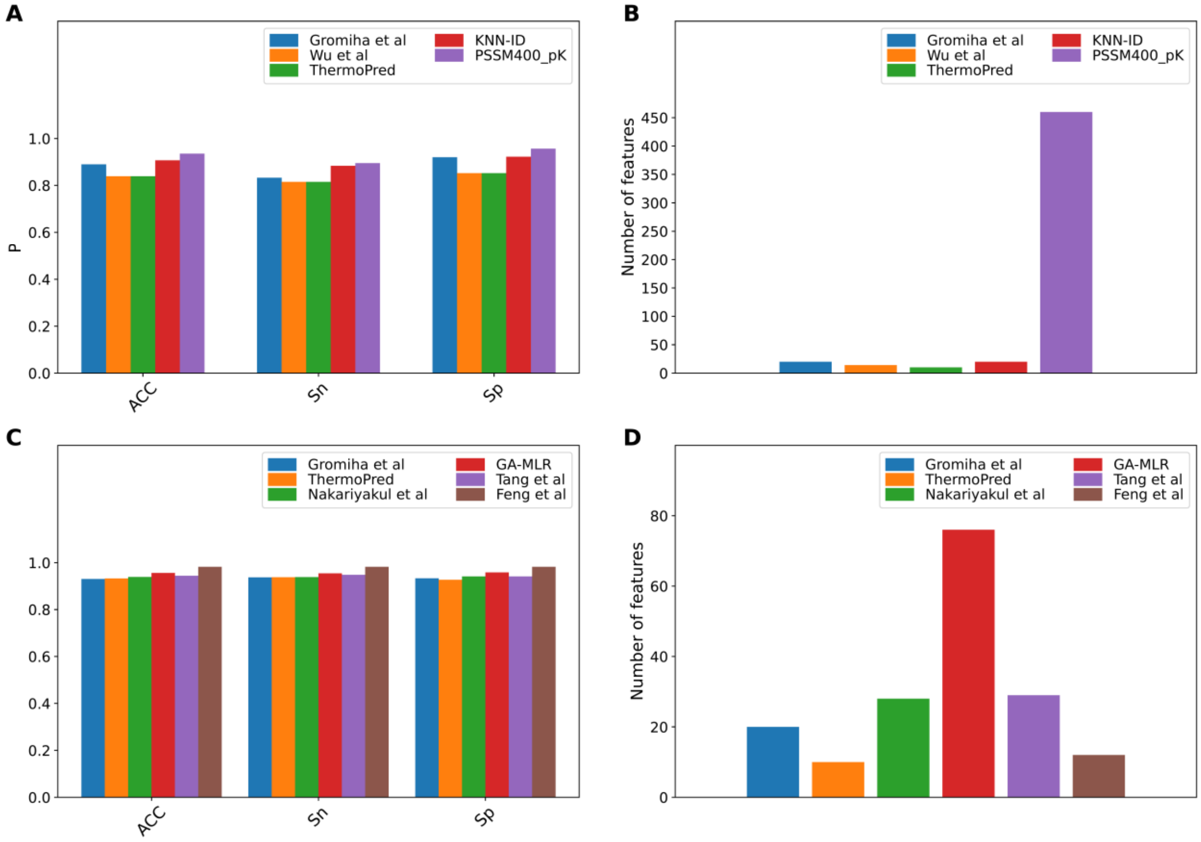


Figure 2: Performance comparison of existing TPP predictors on the Gromiha2007 (**A-B**) and Lin2011 (**C-D**) datasets. (**A, C**) represent the performance in terms of ACC, Sn and Sp. (**B, D**) represent the feature number used in existing TPP predictors.

Received: January 28, 2022, accepted: February 21, 2022, published: March 02, 2022

Table 3: Performance comparison of Gromiha et al.'s method, Wu et al.'s method, ThermoPred, KNN-ID, PSSM400_pK on the Gromiha2007 dataset

Method	Number of features	ACC (%)	Sn (%)	Sp (%)
Gromiha et al.'s method [a,b]	20	89.00	83.30	92.00
Wu et al.'s method [a,c]	14	83.90	81.50	85.20
ThermoPred [a,d]	10	90.80	85.40	93.60
KNN-ID [a,d]	20	90.66	88.37	92.24
PSSM400_pK [a,d]	460	93.53	89.50	95.64

[a] Performance of existing methods were obtained from Fan et al. 2016.
[b] Results based on 5-fold cross-validation
[c] Results based on 10-fold cross-validation
[d] Results based on leave-one-out cross-validation

Table 4: Performance comparison of Gromiha et al.'s method, ThermoPred, Nakariyakul et al.'s method, GA-MLR, Tang et al.'s method and Feng et al.'s method on the Lin2011 dataset

Method	Number of features	ACC (%)	Sn (%)	Sp (%)
Gromiha et al.'s method [a,d]	20	93.00	93.70	93.30
ThermoPred [a,f]	10	93.27	93.77	92.69
Nakariyakul et al.'s method [a, e]	28	93.90	93.80	94.10
GA-MLR [a,d]	76	95.61	95.41	95.84
Tang et al.'s method [a,d]	29	94.40	94.80	94.10
Feng et al.'s method [a,e]	12	98.20	98.20	98.20

[a] Performance of existing methods were obtained from Tang et al., 2017.
[b] Performance of existing methods were obtained from Nakariyakul et al., 2012.
[c] Performance of existing methods were obtained from Wang and Li, 2014.
[d] Results based on 5-fold cross-validation
[e] Results based on 10-fold cross-validation
[f] Results based on leave-one-out cross-validation

From Tables 3-4, the existing TPP predictors were trained and optimized based on the two benchmark datasets and several observations can be made. First, PSSM400_pKa and Feng et al.'s method achieved a better performance compared with the competing TPP predictors on the Gromiha2007 and Lin2011 datasets, respectively. However, their usage and utility is quite limited to experimental scientists. Meanwhile, among several TPP predictors developed using the two benchmark datasets, only ThermoPred was implemented as a web server for the prediction of TPPs. Second, there were two TPP predictors that were evaluated on the two benchmark datasets (i.e., Gromiha et al.'s method and ThermoPred). ThermoPred achieved a competitive performance on the Gromiha2007 and Lin2011 datasets when compared with PSSM400_pKa and Feng et al.'s method, respectively. Altogether, these comparative results indicated that ThermoPred could outperform the competing TPP predictors in terms of both predictive performance and community utility.

Comparative results on the independent test using the Zhang2007 and Charoenkwan2021 datasets

Prediction models having a high cross-validation performance might not perform well on the independent datasets (Charoenkwan et al., 2021a, b; Kabir et al., 2022; Shoombuatong et al., 2017). Thus, in this section, we conducted an independent test to validate and assess the generalization ability of the existing TPP predictors. As can be seen from Table 2, different TPP predictors were validated using different independent datasets.

EXCLI Journal 2022;21:554-570 – ISSN 1611-2156
Received: January 28, 2022, accepted: February 21, 2022, published: March 02, 2022

Moreover, among variant independent datasets, there were three well-known independent datasets derived from the Zhang2006 (76 TPPs and 81 non-TPPs) (Zhang and Fang, 2006), Zhang2007 (382 TPPs and 325 non-TPPs) (Zhang and Fang, 2007) and Charoenkwan2021 (371 TPPs and 371 non-TPPs) (Charoenkwan et al., 2022) datasets. For convenience of discussion, we denote these three independent datasets as Zhang2006TS, Zhang2007TS and Charoenkwan2021TS, respectively.

For the Zhang2006TS dataset, it was first used as the training dataset to develop Zhang et al.'s method. In 2012, Nakariyakul et al., first applied the Zhang2006TS dataset to evaluate their model. In the meanwhile, the Zhang2006TS dataset was used to assess the performance of KNN-ID, GA-MLR and PSSM400_pKa. It should be noted that the Lin2011 (915 TPPs and 793 non-TPPs) and Gromiha2007 (1609 TPPs and 3075 non-TPPs) datasets were utilized to train and optimize Nakariyakul et al.'s method, GA-MLR and KNN-ID and PSSM400_pKa, respectively (Table 2). As can be seen from Table 5, PSSM400_pKa achieved the highest ACC, Sn and Sp of 97.45 %, 97.37 % and 97.53 %, respectively, while the second-best method was KNN-ID in term of ACC. For the Zhang2007TS dataset, it was constructed by Zhang et al (Zhang and Fang, 2007). This dataset was utilized to evaluate the performance of LogitBoost and Gromiha et al.'s method. LogitBoost and Gromiha et al.'s method were trained and optimized using different datasets (Table 2). It could be noticed that LogitBoost outperformed Gromiha et al.'s method in terms of ACC (92.08 %) and Sp (91.70 %) (Table 5). This might be due to the fact that LogitBoost were trained using larger samples. For the last independent dataset, it was constructed by Charoenkwan et al. (2021e). This dataset was utilized to evaluate the performance of ThermoPred and SCMTPP. From Table 5, it could be observed that SCMTPP achieved a very comparable performance to ThermoPred in terms of ACC, Sp and Sn.

Table 5: Performance comparison of LogitBoost, Gromiha et al.'s method, ThermoPred and SCMTPP on three independent datasets

Dataset	Method	Number of features	ACC (%)	Sn (%)	Sp (%)
Zhang2006TS	Nakariyakul et al.'s method [a]	28	86.60	93.40	80.3
	KNN-ID [b]	20	94.20	-	-
	GA-MLR [a]	76	92.99	96.05	90.12
	PSSM400_pKa[c]	460	97.45	97.37	97.53
Zhang2007TS	LogitBoost [d]	20	92.08	91.70	92.52
	Gromiha et al.'s method [e]	20	91.30	87.60	95.70
Charoenkwan2021TS	ThermoPred [f]	10	86.00	93.80	78.20
	SCMTPP [f]	400	86.50	84.90	88.10

[a] Performance of existing methods were obtained from Wang and Li, 2014.
[b] Performance of existing methods were obtained from Zuo et al., 2013.
[c] Performance of existing methods were obtained from Fan et al., 2016.
[d] Performance of existing methods were obtained from Zang and Fang, 2006.
[e] Performance of existing methods were obtained from Gromiha et al.'s method (Gromiha and Suresh, 2008).
[f] Performance of existing methods were obtained from Charoenkwan et al., 2022.

EXCLI Journal 2022;21:554-570 – ISSN 1611-2156
Received: January 28, 2022, accepted: February 21, 2022, published: March 02, 2022

Characterization of TPPs based on sequence information

As mentioned above, there were five out of fourteen existing TPP predictors consisting of Zhang et al.'s method (Zhang and Fang, 2006), LogitBoost (Zhang and Fang, 2007), Wu et al.'s method (Wu et al., 2009), GA-MLR (Wang and Li, 2014) and SCMTPP (Charoenkwan et al., 2022) considered as the computational white-box methods. Amongst these white-box methods developed for the prediction and analysis of TPPs, SCMTPP as introduced by Charoenkwan et al. (2022), outperformed the competing TPP predictors in terms of high interpretability and simplicity. To be specific, SCMTPP was trained and optimized using an up-to-date dataset containing 1853 TPPs and 3233 non-TPPs. By analysis of the propensity scores of twenty amino acids to be TPPs, Charoenkwan et al., reported that the five top-ranked important amino acids to be TPPs were Glu, Lys, Val, Arg and Ile with propensity scores of 510.18, 480.00, 470.75, 464.08 and 435.65, respectively. On the other hand, the five top-ranked important amino acids to be non-TPPs were Gln, Thr, Ala, Asn and Phe with propensity scores of 255.43, 306.00, 323.63, 332.48 and 351.25, respectively. This group also indicated that the top five informative dipeptides to be TPPs consisted of EE, GW, SG, WS and KY with propensity scores of 1000, 979, 956, 952 and 908, respectively, while the top five informative dipeptides to be non-TPPs consisted of AA, LQ, NM, FW and MQ with propensity scores of 0, 11, 27, 41 and 47, respectively. In addition, SCMTPP was applied to determine informative physicochemical properties (PCPs). Charoenkwan et al. (2021e) reported that FUKS010101 (R = 0.616), FUKS010101 (R = 0.523) and FUKS010109 (R = 0.307) were considered as the top three informative PCP used for analyzing TPPs and non-TPPs. Charoenkwan et al.'s analysis showed that the content of hydrophobic amino acids in TPPs was not different from non-TPPs. However, all the results from Charoenkwan et al.'s analysis were derived from primary sequence information, while only selected TPPs and non-TPPs were used to analyze their PCPs. Thus, Charoenkwan et al.'s analysis was limited due to the small size of selected TPPs and non-TPPs used in their studies.

Web server availability and usability

As seen in Table 1, there are a total of 14 sequence-based predictors, but only two of them (ThermoPred and SCMTPP) were implemented as web servers for the prediction of TPPs. ThermoPred is an SVM-based predictor trained with the informative g-gap dipeptides consisted of EE, KE, EI, I-K, I-E, E--K, E--E, K--E, Q--A and E---K derived from A-NOVA approach. ThermoPred is freely available at http://lin-group.cn/server/ThermoPredv1. On the contrary, SCMTPP was constructed using the propensity scores of GGAP (g=0) and a threshold value of 418, where an unknown protein *P* is predicted as TPP if its TPP score is greater than the threshold value, otherwise this protein is predicted as non-TPP. SCMTPP is freely available at http://pmlabstack.pythonanywhere.com/SCMTPP.

PROSPECTIVE STRATEGIES FOR IMPROVING THE PREDICTION PERFORMANCE OF TPPS

In this section, we discuss the advantages and disadvantages of the current TPPs predictors. In addition, we provided future perspectives for the design and development of new computational models for TPP prediction. Hereafter, four crucial aspects for further improving the performance of TPP predictions are discussed and explored.

First, all the existing TPP predictors were trained using sequence-based features. Among several of the sequence-based features employed, ACC was the most frequently used one (Table 1). Numerous previous studies have indicated that sequence-to-vector encodings has been successfully employed for feature extraction in order to facilitate the characterization and analysis of protein, peptide and DNA sequences (Le et al., 2019;

Tahir et al., 2020; Xie et al., 2021). Unlike sequence-based features, sequence-to-vector encodings were able to provide better performance in many cases (Charoenkwan et al., 2021f; Lv et al., 2021; Shah and Ou, 2021; Zulfiqar et al., 2021). To the best of our knowledge, there is no TPP predictor reported that is trained and optimized using sequence-to-vector encodings.

Second, there is no DL-based TPP predictor in existence in this aspect. Meanwhile, with a large number of characterized proteins in recent years, the utility of DL techniques has been reported by numerous studies in biological research (Charoenkwan et al., 2021f; Lv et al., 2021; Shah and Ou, 2021; Xie et al., 2020; Zulfiqar et al., 2021). Specifically, DL-based methods can extract features from protein, peptide and DNA sequences directly by using natural language processing (NLP) technique without the need of feature encodings. To date, DL-based methods are effective and powerful built-in feature extractors. For instance, our group developed BERT4Bitter, which was a bidirectional encoder representation from transformers (BERT)-based model for the identification of bitter peptides (Charoenkwan et al., 2021f). We compared BERT4Bitter with popular ML-based methods developed with ANN, DT, KNN, SVM, ANN, extremely randomized trees (ETree), linear support vector classifier (SVC), logistic regression (LR), naive Bayes (NB), random forest (RF), and extreme gradient boosting (XGB). Specifically, these ML-based methods were trained using five sequence-based features containing AAC, DPC, TC, amino acid index (AAI) and pseudo amino acid composition (PAAC). Remarkably, BERT4Bitter outperformed the comparative ML-based methods in terms of ACC (with an improvement of 2-29 %) and MCC (with improvements of 2-59 %) on the independent dataset. Thus, a DL-based TPP predictor might plausibly achieve improved performance over the fourteen existing TPP predictors.

Third, all the existing TPP predictors were developed by using single ML algorithms to train the model. Thus, their performance is not robust in some cases (Charoenkwan et al., 2021a; Kabir et al., 2022; Liang et al., 2021). To date, there is no ensemble-based TPP predictor in existence in this aspect. There are three popular ensemble learning methods (i.e., majority voting, average probability and stacking strategy). Several studies have indicated that the stacking strategy outperformed the other two ensemble learning methods. Unlike the remaining ensemble learning methods, the stacking strategy can automatically explore different baseline models in order to develop a single stable model. For example, our group proposed a stacking ensemble model, namely StackIL6, for accurately identifying IL-6 inducing peptides (Charoenkwan et al., 2021d). In StackIL6, we employed twelve different feature descriptors and five popular ML algorithms (i.e., ANN, ETree, LR, SVM and RF) to construct variant baseline models for developing the final stacking ensemble model. Our comparative results showed that StackIL6 achieved the highest performance over its baseline models on the training and independent datasets. Altogether, the performance of TPP prediction might be logically increased by applying the ensemble learning strategy (Basith et al., 2022; Charoenkwan et al., 2020; Hasan et al., 2021; Manavalan et al., 2019a, b; Su et al., 2020b; Zhang and Zou, 2020).

Finally, it is well-known that the advantage of a web server is to quickly identify potential TPP candidates from large-scale proteins and provide the prediction without the need to develop an in-house prediction model (Dao et al., 2019; Feng et al., 2019; Lai et al., 2019; Zhu et al., 2019). However, most of these predictors were not developed as web servers, with the exception of ThermoPred and SCMTPP. Although PSSM400_pKa outperformed other TPP predictors, its utility is limited to experimental scientists. Overall, in terms of both predictive performance and community utility, ThermoPred and SCMTPP outperform PSSM400_pKa and other existing TPP predictors.

CONCLUSIONS

In this study, we have conducted empirical comparison and analysis of fourteen existing TPP predictors in terms of multiple perspectives (i.e., feature encoding schemes, feature selection strategies, ML algorithms, evaluation strategies and web server/software usability). We evaluated the existing TPP predictors on the two training and three independent datasets. Our comparative results demonstrated that ThermoPred outperforms other existing TPP predictors in terms of both predictive performance and community utility, while SCMTPP outperforms other existing TPP predictors in terms of high interpretability and simplicity. Although, the existing TPP predictors provide satisfactory prediction performance and promote research progress in this field, there are several issues that need to be addressed. Herein, four crucial aspects for further improving the performance of TPP prediction have been provided as follows: (i) training new models using sequence-to-vector encodings, (ii) using DL-based models, (iii) using an ensemble learning strategy and (vi) developing a web server. We anticipate that this comprehensive review will provide useful insights for researchers in selecting appropriate TPP predictors that are most suitable to deal with their purposes and inspire follow-up research in the future.

Ethical statement

This review paper does not include animal or human experiments.

Conflicts of interest

The authors declare no conflict of interest.

Author contributions statement

WS: Conceptualization, project administration, supervision, investigation, manuscript preparation and revision. PC: Data analysis; data interpretation, investigation and manuscript preparation. NS: Manuscript revision. MMH, MAM and PL: Manuscript preparation. All authors reviewed and approved the manuscript.

Acknowledgments

This work was fully supported by College of Arts, Media and Technology, Chiang Mai University, and partially supported by Chiang Mai University and Mahidol University. In addition, computational resources were supported by Information Technology Service Center (ITSC) of Chiang Mai University.

REFERENCES

Arif M, Ali F, Ahmad S, Kabir M, Ali Z, Hayat M. Pred-BVP-Unb: Fast prediction of bacteriophage Virion proteins using un-biased multi-perspective properties with recursive feature elimination. Genomics. 2020;112:1565-74.

Azadpour M, McKay CM, Smith RL. Estimating confidence intervals for information transfer analysis of confusion matrices. J Acoust Soc Am. 2014;135: EL140-6.

Basith S, Manavalan B, Hwan Shin T, Lee G. Machine intelligence in peptide therapeutics: A next-generation tool for rapid disease screening. Med Res Rev. 2020; 40:1276-314.

Basith S, Lee G, Manavalan B. STALLION: a stacking-based ensemble learning framework for prokaryotic lysine acetylation site prediction. Brief Bioinform. 2022;23(1):bbab376.

Burley SK, Berman HM, Kleywegt GJ, Markley JL, Nakamura H, Velankar S. Protein Data Bank (PDB): The single global macromolecular structure archive. In: Wlodawer A, Dauter Z, Jaskolski M (eds). Protein crystallography: methods and protocols (pp 627-41). New York: Springer, 2017.

Charoenkwan P, Shoombuatong W, Lee H-C, Chaijaruwanich J, Huang H-L, Ho S-Y. SCMCRYS: predicting protein crystallization using an ensemble scoring card method with estimating propensity scores of P-collocated amino acid pairs. PloS One. 2013; 8(9):e72368.

Charoenkwan P, Nantasenamat C, Hasan MM, Shoombuatong W. Meta-iPVP: a sequence-based meta-predictor for improving the prediction of phage virion proteins using effective feature representation. J Computer-Aided Mol Des. 2020;34:1105-16.

Charoenkwan P, Anuwongcharoen N, Nantasenamat C, Hasan M, Shoombuatong W. In silico approaches for the prediction and analysis of antiviral peptides: A review. Curr Pharm Des. 2021a;27:2180-8.

Charoenkwan P, Chiangjong W, Hasan MM, Nantasenamat C, Shoombuatong W. Review and comparative analysis of machine learning-based predictors for predicting and analyzing of anti-angiogenic peptides. Curr Med Chem. 2021b;epub ahead of print.

Charoenkwan P, Chiangjong W, Lee VS, Nantasenamat C, Hasan MM, Shoombuatong W. Improved prediction and characterization of anticancer activities of peptides using a novel flexible scoring card method. Sci Rep. 2021c;1:3017.

Charoenkwan P, Chiangjong W, Nantasenamat C, Hasan MM, Manavalan B, Shoombuatong W. StackIL6: a stacking ensemble model for improving the prediction of IL-6 inducing peptides. Brief Bioinform. 2021d;22(6):bbab172.

Charoenkwan P, Chotpatiwetchkul W, Lee VS, Nantasenamat C, Shoombuatong W. A novel sequence-based predictor for identifying and characterizing thermophilic proteins using estimated propensity scores of dipeptides. Sci Rep. 2021e;11: 23782.

Charoenkwan P, Nantasenamat C, Hasan MM, Manavalan B, Shoombuatong W. BERT4Bitter: a bidirectional encoder representations from transformers (BERT)-based model for improving the prediction of bitter peptides. Bioinformatics. 2021f; 37:2556–62.

Charoenkwan P, Chiangjong W, Nantasenamat C, Moni MA, Lio P, Manavalan B, et al. SCMTHP: A new approach for identifying and characterizing of tumor-homing peptides using estimated propensity scores of amino acids. Pharmaceutics. 2022;14(1):122.

Chen W, Ding H, Feng P, Lin H, Chou K-C. iACP: a sequence-based tool for identifying anticancer peptides. Oncotarget. 2016;7(13):16895.

Dao F-Y, Lv H, Wang F, Feng C-Q, Ding H, Chen W, et al. Identify origin of replication in Saccharomyces cerevisiae using two-step feature selection technique. Bioinformatics. 2019;35:2075-83.

Diaz JE, Lin C-S, Kunishiro K, Feld BK, Avrantinis SK, Bronson J, et al. Computational design and selections for an engineered, thermostable terpene synthase. Protein Sci. 2011;20:1597-606.

Fan G-L, Liu Y-L, Wang H. Identification of thermophilic proteins by incorporating evolutionary and acid dissociation information into Chou's general pseudo amino acid composition. J Theor Biol. 2016;407:138-42.

Feng C-Q, Zhang Z-Y, Zhu X-J, Lin Y, Chen W, Tang H, et al. iTerm-PseKNC: a sequence-based tool for predicting bacterial transcriptional terminators. Bioinformatics. 2019;35:1469-77.

Feng C, Ma Z, Yang D, Li X, Zhang J, Li Y. A method for prediction of thermophilic protein based on reduced amino acids and mixed features. Front Bioeng Biotechnol. 2020;8:285.

Gaucher EA, Govindarajan S, Ganesh OK. Palaeotemperature trend for Precambrian life inferred from resurrected proteins. Nature. 2008;451(7179):704-7.

Gromiha MM. Protein bioinformatics. Cambridge, MA: Academic Press, 2010.

Gromiha MM, Suresh MX. Discrimination of mesophilic and thermophilic proteins using machine learning algorithms. Proteins. 2008;70:1274-9.

Gromiha MM, Oobatake M, Sarai A. Important amino acid properties for enhanced thermostability from mesophilic to thermophilic proteins. Biophys Chem. 1999;82:51-67.

Gromiha MM, Nagarajan R. Computational approaches for predicting the binding sites and understanding the recognition mechanism of protein–DNA complexes. In: Donev R (ed). Advances in protein chemistry and structural biology (pp 65-99). Cambridge, MA: Academic Press, 2013.

Gromiha MM, Nagarajan R, Selvaraj S. Protein structural bioinformatics: an overview. In: Ranganathan S, Gribskov M, Nakai K, Schönbach C (eds). Encyclopedia of bioinformatics and computational biology (pp 445-59. Cambridge, MA: Academic Press, 2019.

Habbeche A, Saoudi B, Jaouadi B, Haberra S, Kerouaz B, Boudelaa M, et al. Purification and biochemical characterization of a detergent-stable keratinase from a newly thermophilic actinomycete Actinomadura keratinilytica strain Cpt29 isolated from poultry compost. J Biosci Bioeng. 2014;117:413-21.

Haki GD, Rakshit SK. Developments in industrially important thermostable enzymes: a review. Bioresour Technol. 2003;89(1):17-34.

Hasan MM, Schaduangrat N, Basith S, Lee G, Shoombuatong W, Manavalan B. HLPpred-Fuse: improved and robust prediction of hemolytic peptide and its activity by fusing multiple feature representation. Bioinformatics. 2020;36:3350-6.

Hasan MM, Basith S, Khatun MS, Lee G, Manavalan B, Kurata H. Meta-i6mA: an interspecies predictor for identifying DNA N 6-methyladenine sites of plant genomes by exploiting informative features in an integrative machine-learning framework. Brief Bioinform. 2021;22(3):bbaa202.

Huang H-L, Charoenkwan P, Kao T-F, Lee H-C, Chang F-L, Huang W-L, et al. Prediction and analysis of protein solubility using a novel scoring card method with dipeptide composition. BMC Bioinformatics. 2012a;13(Suppl 17):S3.

Huang SY, Zhang YH, Zhong JJ. A thermostable recombinant transaldolase with high activity over a broad pH range. Appl Microbiol Biotechnol. 2012; 93b:2403-10.

Kabir M, Nantasenamat C, Kanthawong S, Charoenkwan P, Shoombuatong W. Large-scale comparative review and assessment of computational methods for phage virion proteins identification. EXCLI J. 2022; 21:11-29.

Kurgan L, Razib AA, Aghakhani S, Dick S, Mizianty M, Jahandideh S. CRYSTALP2: sequence-based protein crystallization propensity prediction. BMC Struct Biol. 2009;9(1):1-14.

Lai H-Y, Zhang Z-Y, Su Z-D, Su W, Ding H, Chen W, et al. iProEP: a computational predictor for predicting promoter. Mol Ther Nucleic Acids. 2019;17:337-46.

Le NQK, Yapp EKY, Ho Q-T, Nagasundaram N, Ou Y-Y, Yeh H-Y. iEnhancer-5Step: Identifying enhancers using hidden information of DNA sequences via Chou's 5-step rule and word embedding. Anal Biochem. 2019;571:53-61.

Li J, Zhu P, Zou Q. Prediction of thermophilic proteins using voting algorithm. In: Rojas I, Valenzuela O, Rojas F, Ortuño F (eds). Bioinformatics and biomedical engineering. IWBBIO 2019 (pp 195-203). Cham: Springer, 2019. (Lecture Notes in Computer Science, Vol. 11465).

Liang X, Li F, Chen J, Li J, Wu H, Li S, et al. Large-scale comparative review and assessment of computational methods for anti-cancer peptide identification. Brief Bioinform. 2021;22(4):bbaa312.

Lin H, Chen W. Prediction of thermophilic proteins using feature selection technique. J Microbiol Meth. 2011;84:67-70.

Lv H, Dao F-Y, Zulfiqar H, Lin H. DeepIPs: comprehensive assessment and computational identification of phosphorylation sites of SARS-CoV-2 infection using a deep learning-based approach. Brief Bioinform. 2021;22(6):bbab244.

Manavalan B, Lee J. SVMQA: support–vector-machine-based protein single-model quality assessment. Bioinformatics. 2017;33:2496-503.

Manavalan B, Basith S, Shin TH, Wei L, Lee G. mAHTPred: a sequence-based meta-predictor for improving the prediction of anti-hypertensive peptides using effective feature representation. Bioinformatics. 2019a;35:2757-65.

Manavalan B, Basith S, Shin TH, Wei L, Lee G. Meta-4mCpred: A sequence-based meta-predictor for accurate DNA 4mC site prediction using effective feature representation. Mol Ther Nucleic Acids. 2019b;16: 733-44.

Nakariyakul S, Liu Z-P, Chen L. Detecting thermophilic proteins through selecting amino acid and dipeptide composition features. Amino Acids. 2012;42:1947-53.

Narasimhan D, Nance MR, Gao D, Ko M-C, Macdonald J, Tamburi P, et al. Structural analysis of thermostabilizing mutations of cocaine esterase. Protein Eng Des Sel. 2010;23:537-47.

Pica A, Graziano G. Shedding light on the extra thermal stability of thermophilic proteins. Biopolymers. 2016;105:856-63.

Rao B, Zhou C, Zhang G, Su R, Wei L. ACPred-Fuse: fusing multi-view information improves the prediction of anticancer peptides. Brief Bioinform. 2020;21:1846-55.

Rodriguez E, Mullaney EJ, Lei XG. Expression of the Aspergillus fumigatus phytase gene in Pichia pastoris and characterization of the recombinant enzyme. Biochem Biophys Res Commun. 2000;268:373-8.

Shah SMA, Ou Y-Y. TRP-BERT: Discrimination of transient receptor potential (TRP) channels using contextual representations from deep bidirectional transformer based on BERT. Comput Biol Med. 2021; 137:104821.

Shoombuatong W, Prathipati P, Owasirikul W, Worachartcheewan A, Simeon S, Anuwongcharoen N, et al. Towards the revival of interpretable QSAR models. In: Roy K (ed). Advances in QSAR modeling (pp 3-55). Cham: Springer, 2017 (Challenges and Advances in Computational Chemistry and Physics, Vol. 24).

Su R, Hu J, Zou Q, Manavalan B, Wei L. Empirical comparison and analysis of web-based cell-penetrating peptide prediction tools. Brief Bioinform. 2020a;21: 408-20.

Su R, Liu X, Xiao G, Wei L. Meta-GDBP: a high-level stacked regression model to improve anticancer drug response prediction. Brief Bioinform. 2020b;21:996-1005.

Received: January 28, 2022, accepted: February 21, 2022, published: March 02, 2022

Tahir M, Hayat M, Chong KT. Prediction of N6-methyladenosine sites using convolution neural network model based on distributed feature representations. Neural Networks. 2020;129:385-91.

Tang H, Cao R-Z, Wang W, Liu T-S, Wang L-M, He C-M. A two-step discriminated method to identify thermophilic proteins. Int J Biomath. 2017;10(4): 1750050.

Vieille C, Zeikus GJ. Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability. Microbiol Mol Biol Rev. 2001;65(1): 1-43.

Wang D, Yang L, Fu Z, Xia J. Prediction of thermophilic protein with pseudo amino acid composition: an approach from combined feature selection and reduction. Protein Pept Lett. 2011;18: 684-9.

Wang L, Li C. Optimal subset selection of primary sequence features using the genetic algorithm for thermophilic proteins identification. Biotechnol Lett. 2014;36: 1963-9.

Wu L-C, Lee J-X, Huang H-D, Liu B-J, Horng J-T. An expert system to predict protein thermostability using decision tree. Expert Syst Applicat. 2009;36:9007-14.

Xie R, Li J, Wang J, Dai W, Leier A, Marquez-Lago TT, et al. DeepVF: a deep learning-based hybrid framework for identifying virulence factors using the stacking strategy. Brief Bioinform. 2021;22(3): bbaa125.

Xu H, Shen D, Wu XQ, Liu ZW, Yang QH. Characterization of a mutant glucose isomerase from Thermoanaerobacterium saccharolyticum. J Ind Microbiol Biotechnol. 2014;41:1581-9.

Zhang G, Fang B. Discrimination of thermophilic and mesophilic proteins via pattern recognition methods. Process Biochem. 2006;41(3):552-6.

Zhang G, Fang B. LogitBoost classifier for discriminating thermophilic and mesophilic proteins. J Biotechnol. 2007;127:417-24.

Zhang YP, Zou Q. PPTPP: a novel therapeutic peptide prediction method using physicochemical property encoding and adaptive feature representation learning. Bioinformatics. 2020;36:3982-7.

Zhu X-J, Feng C-Q, Lai H-Y, Chen W, Hao L. Predicting protein structural classes for low-similarity sequences by evaluating different features. Knowledge-Based Sys. 2019;163:787-93.

Zulfiqar H, Sun Z-J, Huang Q-L, Yuan S-S, Lv H, Dao F-Y, et al. Deep-4mCW2V: a sequence-based predictor to identify N4-methylcytosine sites in Escherichia coli. Methods. 2021;epub ahead of print.

Zuo Y-C, Chen W, Fan G-L, Li Q-Z. A similarity distance of diversity measure for discriminating mesophilic and thermophilic proteins. Amino Acids. 2013;44: 573-80.