RESEARCH ARTICLE

A computational method for prediction of xylanase enzymes activity in strains of *Bacillus subtilis* based on pseudo amino acid composition features

Shohreh Ariaeenejad[1]*, **Maryam Mousivand**[2], **Parinaz Moradi Dezfouli**[2], **Maryam Hashemi**[2], **Kaveh Kavousi**[3], **Ghasem Hosseini Salekdeh**[1]

1 Department of Systems Biology, Agricultural Biotechnology Research Institute of Iran (ABRII), Agricultural Research Education and Extension Organization (AREO), Karaj, Iran, **2** Department of Microbial Biotechnology, Agricultural Biotechnology Research Institute of Iran (ABRII), Agricultural Research Education and Extension Organization (AREO), Karaj, Iran, **3** Institute of Biochemistry and Biophysics (IBB), University of Tehran, Tehran, Iran

* Sh.ariaee@abrii.ac.ir, shariaee@gmail.com

Citation: Ariaeenejad S, Mousivand M, Moradi Dezfouli P, Hashemi M, Kavousi K, Hosseini Salekdeh G (2018) A computational method for prediction of xylanase enzymes activity in strains of *Bacillus subtilis* based on pseudo amino acid composition features. PLoS ONE 13(10): e0205796. https://doi.org/10.1371/journal.pone.0205796

Editor: Jinn-Moon Yang, National Chiao Tung University College of Biological Science and Technology, TAIWAN

Received: August 22, 2017

Accepted: October 2, 2018

Published: October 22, 2018

Data Availability Statement: All relevant data and amino acid sequences as well as their GenBank accession numbers are within the paper and its Supporting Information files.

Funding: The authors received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Xylanases are hydrolytic enzymes which based on physicochemical properties, structure, mode of action and substrate specificities are classified into various glycoside hydrolase (GH) families. The purpose of this study is to show that the activity of the members of the xylanase family in the specified pH and temperature conditions can be computationally predicted. The proposed computational regression model was trained and tested with the Pseudo Amino Acid Composition (PseAAC) features extracted solely from the amino acid sequences of enzymes. The xylanases with experimentally determined activities were used as the training dataset to adjust the model parameters. To develop the model, 41 strains of *Bacillus subtilis* isolated from field soil were screened. From them, 28 strains with the highest halo diameter were selected for further studies. The performance of the model for prediction of xylanase activity was evaluated in three different temperature and pH conditions using stratified cross-validation and jackknife methods. The trained model can be used for determining the activity of newly found xylanases in the specified condition. Such computational models help to scale down the experimental costs and save time by identifying enzymes with appropriate activity for scientific and industrial usage. Our methodology for activity prediction of xylanase enzymes can be potentially applied to the members of the other enzyme families. The availability of sufficient experimental data in specified pH and temperature conditions is a prerequisite for training the learning model and to achieve high accuracy.

Introduction

After cellulose, xylan is the second most abundant polysaccharide in nature which is mostly found in the plant cell wall and accounts for a large part of plants biomass. Xylan can be depolymerized using xylanase enzymes, an important family of hydrolases.

The glycosyl hydrolases (GHs) are a very large family of enzymes which hydrolyze the glycosidic bond between carbohydrates as well as between a carbohydrate and a noncarbohydrate moiety to form heteropolysaccharides. The classification of GH enzymes into subfamilies is mainly based on amino-acid sequence similarities as proposed in [1–7].

Endo-xylanases with somewhat different sequences are found in various GH families because of the sequence-based classification of GH enzymes and despite similar structures and conserved folding [3–9].

Xylanases (EC3.2.1.X) are among important constituent subfamilies of GH enzymes. While xylanase isoenzymes show different specificities, they have synergistic effect on the hydrolysis of xylan [10]. Heteroxylan backbone is composed of glycoside linkages. For cleaving these bonds, the interaction of some cleavage enzymes for both main and side chains is required.

Endo-β-1,4-xylanases (EC 3.2.1.8), β-1,4 xylosidases (EC 3.2.1.37), and exoxylanases are examples of enzymes with the capability of cleaving main-chain glycosyl groups [7]. Most xylanases extracted from microbial communities are single-subunit enzymes [9].

There have been a lot of works to achieve highly active xylanases suitable for various applications in specified conditions [11–13]. A comprehensive review covered those approaches and offered a procedure for cloning of recombinant xylanase enzymes with thermostability and alkaline stability [14]. There are many computational approaches for predicting the enzyme activity from its tertiary structure [15,16], but the prediction of the activity of an enzyme based on its sequence is not a straightforward task. The members of a specific enzyme family, e.g., xylanases, have very similar sequences with high sequence identities, but very different activity levels in similar conditions. This property makes it very hard to predict the activity only from the sequence. The purpose of the proposed computational method is to predict the activity of enzymes from xylanase family based on limited experimental studies.

Different *Bacillus subtilis* strains capable of xylanase production have been hitherto isolated from natural resources [17–23]. In this study, 41 strains of *Bacillus subtilis* were isolated from gardens and farms based on their ability to produce xylanase enzyme.

For these strains, the xylanase activity determination experiment was done. Using trained computational models, the halo zone diameter in screening plates as well as enzyme activities, could be predicted based on Pseudo Amino Acid Composition (PseAAC) features that were extracted from xylanase amino acid sequences. This makes it possible to predict the bacterial halo diameter and enzyme activity in specified condition without doing screening and activity measurement experiments.

The main reason for choosing PseAAC feature vectors as representative of xylanase enzymes in activity prediction task was the fact that PseAAC features have been vastly used in computational biology for prediction of different properties of proteins and nucleic acid sequences since 2001 [24–58]. Some of its recent applications are related to RNA and DNA sequence analysis fields. Pseudo k-tuple nucleotide compositions (PseKNC) were exploited to identify enhancers and their strength in a two-layer architecture and since 2015 it has been accessible via iEnhancer-2L web server [34]. In 2016, two ensemble learning methods were introduced. The iDHS-EL is a web server for identifying DNase I hypersensitive sites which fuses three different modes of pseudo nucleotide composition [33]. Also, the iRSpot-EL fuses different modes of PseKNC plus mode of dinucleotide-based auto-cross covariance for identifying DNA recombination spots [59]. One of the most recent studies in 2017 introduces

2L-piRNA, a two-layer ensemble classification system, for identifying Piwi-Interacting RNAs and their function using PseKNC [60].

Among the important factors in industrial processes are pH and temperature on which chemical and enzymatic stability depend. Therefore, choosing the right enzyme to optimize catalyzing a specific reaction is not straightforward [61].

Many attempts have been made for engineering thermostable microbial xylanases for optimizing their activity in industrial processes through advanced biotechnological approaches including enzyme immobilization methods, gene editing and docking [62–64]. Despite the above mentioned studies, there is still no computational framework to predict the enzyme specific activity in the specified condition. The proposed approach can facilitate this complex process using statistical learning methods. Moreover, this method can be extensively used for screening the activities of enzymes extracted from metagenomic data.

Materials and methods

Experimental data

Bacterial strains and culture condition. About 90 *Bacillus subtilis* isolates were obtained from Microbial Culture Collection established in the Agricultural Biotechnology Research Institute of Iran (ABRII). The strains were grown in NBY medium (Nutrient Broth: 8g / L, K_2HPO_4: 1g / L, Yeast Extract: 1g / L, KH_2PO_4: 0.25g / L, Glucose: 2g / l, $MgSO_4$ (1M): 1ml / L and Agar: 18g / L) and incubated at 28˚C for 48 h.

Screening of xylanase producing bacterial isolates. Bacterial isolates were grown on XC agar medium containing 10 g/L oat-spelt xylan, 5 g/L peptone, 1 g/L yeast extract, 4 g/L K_2HPO_4, 1 g/L $MgSO_4.7H_2O$, 0.2 g/L KCl, 0.02 g/L $FeSO_4.7H_2O$, agar 15 g/L, pH 7.0. The plates were incubated at 28˚C for 48h. Xylanse producing bacteria exhibited a clear zone around their colony as a qualitative index for xylanase productivity potential.

Enzyme production. For crude enzyme production, 200μl of overnight-grown bacterial culture in nutrient broth (OD600nm = 0.5) was transferred into 10 ml enzyme medium and shaked at 28˚C for 48h. The enzyme medium contained xylan: 12 g/L, Meat Extract: 3 g/L, Yeast Extract: 4 g/L, $CaCl_2.H_2O$: 0.5 g/L, $MgSO_4.7H_2O$: 0.3 g/L and K_2HPO_4: 1 g/L and pH was adjusted to 7.0. The fermented culture medium was centrifuged at 10,000 rpm for 10 min at 4˚C and the supernatant was stored at -20˚C for xylanase assay.

Xylanase assay. Xylanase activity was assayed by measuring the formation of reducing sugar by the dinitrosalicylic acid (DNS) method [65]. The reaction mixture containing 100 μl of crude enzyme and 300 μl 1%soluble xylan(sigma) in 50 mM phosphate or citrate buffer at desired pH. After 20 min, the 600 μl DNS reagent was added to the mixture and boiled at 100˚C for 15 min. The xylanase was assayed at three different conditions including temperature = 60˚C and pH = 4.6, temperature = 26˚C and pH = 4.6 and temperature = 26˚C and pH = 6.9. Absorbance was measured at 540 nm against a reagent blank. A series of xylose dilutions were used as standards to calculate the quantity of reduced sugar. One unit (U) of xylanase activity was defined as the amount of enzyme needed to generate 1 μmol of reduced sugar per minute under the assay conditions.

Collected dataset. The xylanases were extracted from 41 different strains of *Bacillus subtilis*. Their GenBank accession numbers and the associated strain codes are demonstrated in Table 1. Also, Their amino acid sequences are provided in supplementary S2 Table. The diameter of bacterial halos was measured. Among 41 xylanases mentioned in Table 1, 28 enzymes were selected for determining their activities in different conditions of temperature and pH.

Table 1. 41 different xylanase enzymes were selected for experimental and computational studies. The GenBank Accession No., and its relevant strain code, for each sequence are included. The diameter of halos produced in the screening plates is also included for enzymes with high and medium halos surface. The last three columns shows the activities measured for 28 selected xylanase enzymes in three different pH and temperature conditions.

					Activity (IU ml^{-1})		
No.	***GenBank Accession No.***	***Strain***	***Halo zone diameter (mm)***	***Real Class***	***pH = 4 T = 60°C***	***pH = 4 T = 26°C***	***pH = 6 T = 26°C***
1	AGO02713	a14h	4.6	M	400	280	320
2	AGO02715	d16d	5.7	H	295	80	300
3	AGO02724	d19d	5	M	490	360	370
4	AGO63347	d3d	5	M	60	120	150
5	AGO63342	h11h	5	M	590	380	205
6	AGO02730	h13f	4.2	M	170	200	150
7	AGS78259	h13h	4.1	M	490	370	320
8	AGO63354	h14d	6.5	H	740	120	170
9	AGO63351	h14h	5	M	810	130	330
10	AGO63345	h16h	4.8	M	330	205	150
11	AGO63356	k2b	7	H	670	230	250
12	AGO02722	k32l	5	M	280	290	150
13	AGO02727	k33l	5	M	440	660	330
14	AGO63344	k36p88	5	M	400	230	180
15	AGO63350	k40b	6	H	610	320	200
16	AGO02714	k43l	5	M	220	180	50
17	AGO02728	k46b	4	M	510	60	210
18	AGO02725	s6a	5.8	H	710	420	40
19	AGO02721	s7e	6.5	H	890	420	780
20	AGO02726	S7h	5	M	370	350	280
21	AGO97103	t27b	4.3	M	530	280	210
22	AGO02717	t28d	5	M	525	170	150
23	AGO63355	t31d	4	M	5	40	50
24	AGO63353	t34b	4.5	M	210	0	120
25	AGO02716	t37a	8	H	670	280	590
26	AGO02718	t41a	5	H	505	310	390
27	AGO02729	W	4.5	M	410	110	125
28	AGO63357	Y	4.5	M	690	390	260
29	AGO02712	b16b	3	L			
30	AGO02719	s2f	2.9	L			
31	AGO02720	s2h	2.5	L			
32	AGO02732	a10d	3.5	M			
33	AGO02723	d3b	2.5	L			
34	AGO02731	b5d	3	L			
35	AGO02733	s3d	2	L			
36	AGO63358	b9h	2.7	L			
37	AGO63360	s5d	3	L			
38	AGO97104	h13d	3	L			
39	AGO02734	S1d	3.8	M			
40	AGO63359	k27k88	3.5	M			
41	AGO63349	b11h	3.5	M			

All cloned xylanase gene sequences belongs to the CAZy GH family 11 according to the Expert Protein Analysis System (ExPASy) PROSITE.

The xylanases in rows 38–41 are excluded because they showed very low activities in all three different conditions of temperature and pH. By experimentally determining the activities for 28 sequences in three conditions, they were used as the material for building and validating a regression model to predict the activity of the xylanase enzymes. The model was validated using stratified k-fold cross validation and jackknife methods.

https://doi.org/10.1371/journal.pone.0205796.t001

Computational analysis

Feature extraction. From the machine learning perspective and, for computational prediction of enzymes activity solely from the sequence, the first step is extracting informative feature vectors based on the amino acid sequence of enzymes. These vectors are considered as the identity profile for each member of the enzyme family. It is expected that using these discriminative profiles and employing powerful computational methods, the activity level of novel enzyme sequences can be estimated. For learning the predictive model for a specific enzyme family, it is necessary to experimentally obtain the enzymatic activities for a limited number of enzyme sequences as training data. One of the well-known sequence based features which has been used in many computational tasks is the amphiphilic Pseudo Amino Acid Composition (PseAAC).

The concept of PseAAC was proposed by Chou [25]. Since then, the concept of PseAAC has penetrated into almost all the fields of computational proteomics [26–30,58]. Encouraged by the successes of introducing the PseAAC approach into computational proteomics, a novel feature vector, called 'pseudo K-tuple nucleotide composition'(PseKNC) [31,32], was developed to represent DNA sequence samples to improve the quality of predicting the elements [33–37,39,40,57,66]. Some soft packages or web servers were established to produce the PseKNC [41–43]. The Pse-in-One is a web server with the ability of generating totally 28 different modes of pseudo components for DNA, RNA, and protein sequences [43]. Also, the Pse-Analysis is a Python package freely accessible at http://bioinformatics.hitsz.edu.cn/Pse-Analysis/ [67]. It provides an automated pipeline including feature extraction from samples and parameter selection, training and validating the model, and evaluating the quality of prediction.

The method of calculating the PseAAC vectors from the amino acid sequence is described in details in [68] and [24].

Suppose an enzyme E, with a sequence of L amino acid residues:

$$E = E_1E_2E_3 \ldots E_L \tag{1}$$

In which E_i (i = 1,2,. . .,L) denotes the residue at chain position i. The hydrophobicity or hydrophilicity of amino acids plays important role in enzyme structure and hence its function [44]. Therefore, these indices are strong candidates to reflect the function and activity of enzyme sequences. The following equations reflect the sequence order effect of an enzyme in its activity and functionality:

$$\begin{cases} \tau_{2k-m} = \dfrac{1}{L-k}\sum_{i=1}^{L-k} H^{m}_{i,i+k} \\ k = 1, 2, \ldots, \lambda;\ \lambda < L \\ m = 0\ or\ 1 \end{cases} \tag{2}$$

In above equations, τ_{2k-1} and τ_{2k} are called the k^{th}-tier correlation factors and $H^1_{i,i+k}$ and $H^2_{i,i+k}$ are the hydrophobicity (m = 0) and hydrophilicity (m = 1) correlation functions respectively.

τ_{2-1} and τ_{2k} reflect the sequence–order amphiphilic correlation between all the k^{th} most contiguous residues along the enzyme sequence. For example, τ_5 and τ_6 are the 3rd-tier (k = 3) correlation factors that shows the sequence-order correlation between all the 3rd most contiguous residues in the sequence (Fig 1).

We used the PseAAC, a web server which is designed to generate PseAAC features from protein sequences [46] (http://www.csbio.sjtu.edu.cn/bioinf/PseAAC/).

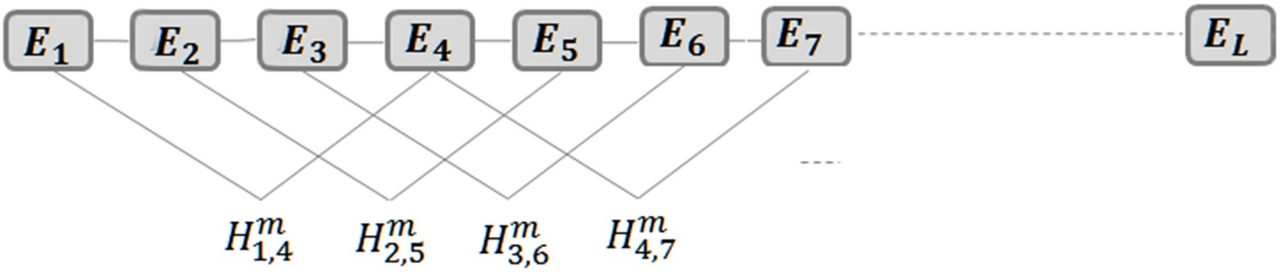

Fig 1. The amphiphilic coupling between all the third most contiguous amino acids. The values 0 and 1 for *m* represents the correlations via hydrophobicity and hydrophilicity indices.

https://doi.org/10.1371/journal.pone.0205796.g001

For each enzyme sample E, we have an augmented vector to represent it:

$$E = \begin{bmatrix} e_1 \\ \vdots \\ e_{20} \\ e_{21} \\ \vdots \\ e_{20+\lambda} \\ e_{20+\lambda+1} \\ \vdots \\ e_{20+2\lambda} \end{bmatrix} \tag{3}$$

The elements of *E* are defined follows:

$$e_k = \begin{cases} \dfrac{f_k}{\sum_{i=1}^{20} f_i + w\sum_{j=1}^{2\lambda} \tau_j}; (1 \leq k \leq 20) \\ \dfrac{w\tau_k}{\sum_{i=1}^{20} f_i + w\sum_{j=1}^{2\lambda} \tau_j}; (21 \leq k \leq 20 + 2\lambda) \end{cases} \tag{4}$$

Different values for λ produce different feature vectors for enzymes. In this study, due to slightly better performance, $\lambda = 7$ is used for generating feature vectors.

The feature values extracted from studied xylanase sequences are tabulated in supplementary S1 Table.

Building classification models. For constructing and learning a model to predict the bacterial halo diameters without the need for experimental works, 41 collected strains were cultured in selective environments and the halo diameters were measured. These results and PseAAC features obtained from respective xylanase sequence were used as the training and testing datasets for classifiers.

The Naïve Bayes, SVM (Support Vector Machine), K-Nearest Neighbor (KNN) (with K = 1) and random forest classifiers were used classify the diameter of halos. For KNN classifier with uniform weights and Euclidean distance, different values for K (from 1 to 15) were considered and for K = 1 the best performance was obtained.

SVM classifier with linear kernel, and random forest classifier with 30 trees were employed. The target class has been selected based on the majority vote from the individually trained trees in the forest. These classifiers were also used in many previous studies including virion protein prediction [48] DNA/RNA modified site identification [47,69], membrane transporter prediction [49] and the origin of replication prediction [39].

To compare the methods, the Area Under Curve (AUC) of ROC curve, Classification Accuracy (CA), F1, precision and recall measures obtained from classification methods were used.

Building regression models. The statistical process of estimating the relationships between a dependent variable and one or more independent variables is called regression. In fact, the conditional expectation of dependent variable is estimated given the independent variables, or predictors. In this study, the enzyme activity in a fixed pH and temperature was estimated based on PseAAC features.

We used different regression methods to determine the activity of xylanases with slightly different sequences using PseAAC vectors in three different pH-temperature conditions.

SVM, KNN and random forest regression algorithms were used to build a proper regression model for xylanase activities in different conditions. Also, boosting regression trees using Adaboost algorithm were examined. For SVM regressor, linear kernel was employed. For KNN, uniform weights and Euclidean distance were considered and different values for K (from 1 to 15) were tested. The best performance was achieved for k = 5. In Adaboost, 50 regression trees as the base regressor machine were fused. For each regression tree, at least two instances for each leaf and 5 instances for internal nodes with the maximum depth of 100 were considered. In the random forest regressor, 10 trees were generated and for each tree maximal tree depth and an unlimited number of considered features were used. The Mean Squared Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and R2 measures were calculated and used for comparing the results.

Implementation and validation of computational models. In this study, Orange was used for performing classification, regression and validation operations as an open source datamining and machine learning toolbox implemented in python [70]. Free access to python codes of Orange makes it possible to use it in the future development of web applications for similar studies.

K-fold cross-validation test, sub-sampling test, independent dataset test and jackknife cross-validation test are four kinds of strategies in statistical learning which have been widely used to examine the performance of a prediction model [71–73]. Because the jackknife test can achieve unique outcomes [74], it has been widely used in Bioinformatics [75–79]. However, the jackknife cross-validation is more time-consuming. In this study, the 10-fold cross-validation as well as the jackknife method were used to investigate the performance of the prediction models.

Results

Xylanase assay

Experimental screening resulted in isolation and identification of 41 isolates producing xylanase enzyme. Approximately, 28 xylanase producing strains (68%) had clear zones larger than 35 mm and selected for xylanase assay at three different conditions.

The halos with a diameter less than 3.5mm are assigned to class Low (L), between 3.5mm and 5.5mm are assigned to class Medium (M), and larger than 5.5mm are assigned to class High (H). Among these xylanases, 28 sequences related to strains with Halo Diameter (HD) greater than 3.5mm (from M or H classes) were selected for further analysis.

Classification results

Applying Random Forest, Naïve Bayes, SVM, and KNN on 41 sequences listed in Table 1 for classifying the area of bacterial halos from respective expressed xylanase enzymes showed that the diameter and therefore the area of these halos could be classified with high accuracy in one of the three categories L, M, or H. Table 2 shows the results. The AUC, CA, F1, precision and

Table 2. Results from three different classifiers.

Model	*AUC*		*CA*		*F1*		*Precision*		*Recall*	
	10 fold	**Jackknife**	**10 fold**	**Jackknife**	**10 fold**	**Jackknife**	**10 fold**	**Jackknife**	**10 fold**	**Jackknife**
Random Forest	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
SVM	0.857	0.854	0.854	0.829	0.856	0.831	0.861	0.834	0.854	0.829
KNN (K = 1)	0.964	0.910	0.951	0.902	0.951	0.900	0.954	0.901	0.951	0.902
Naïve Bayes	0.688	0.636	0.317	0.610	0.307	0.604	0.608	0.600	0.317	0.610

https://doi.org/10.1371/journal.pone.0205796.t002

recall measures for different models are reported in Table 2. The results were validated by 10-fold cross validation and jackknife methods.

According to the results in Table 2, the Random Forest classifier outperformed other models.

Regression results

Using SVM, KNN, Adaboost and random forest regression algorithms, predictive regression models were built for predicting the xylanase activities in three specific temperatures and pH conditions. The experimental data in Table 2 were used for tuning the parameters of these models. The results are demonstrated in Fig 2 and Table 3. Fig 2, shows the experimentally measured activities vs. predicted values by all four regression models for all 28 strains. Table 3 summarizes the performance measures for regression models. Based on the results, the SVM regressor showed the best performance. As it can be seen in Fig 2, except the s7e and t31d in all three conditions, and t34b in part (a) almost for all other strains the activity of produced enzyme has been accurately predicted by at least one of predictors. Despite the overall better performance of SVM, the Random Forest regressor showed better results.

Discussion and conclusion

Understanding the properties of amino acid sequences from their primary structure is one of the main challenges in computational biology.

The rapid growth in the number of enzymes discovered from high-throughput sequencing generates a wealth of data. However, a major challenge is the functional assignment and activity prediction for many newly found enzymes with no or limited experimental data. The activity of an enzyme in a specified condition is a very important factor that can affect the rate of the underlying reaction.

It is worth noting that both enzyme molecular function prediction and enzyme specific activity prediction are important and challenging subjects which should not be confused with each other.

Enzyme molecular function prediction refers to identifying the biochemical reactions that an enzyme can catalyze and these functions are manually classified by the Enzyme Commission[80]. Several in-silico and experimental methods have been developed for this purpose, many of which are based on the identification of target substrates for the enzyme active site(s) [50].

However, for members of an enzyme family with similar molecular function, the level of catalytic specific activity can be very different for a given condition of temperature, pH and the presence of inhibitory factors. Establishing a computational framework for in-silico prediction of the specific activities for the members of an enzyme family only from their amino acid sequence, and for a given condition, is the main novelty of this research. Building a learning model with high generalization power needs adequate training samples. In this field, we need

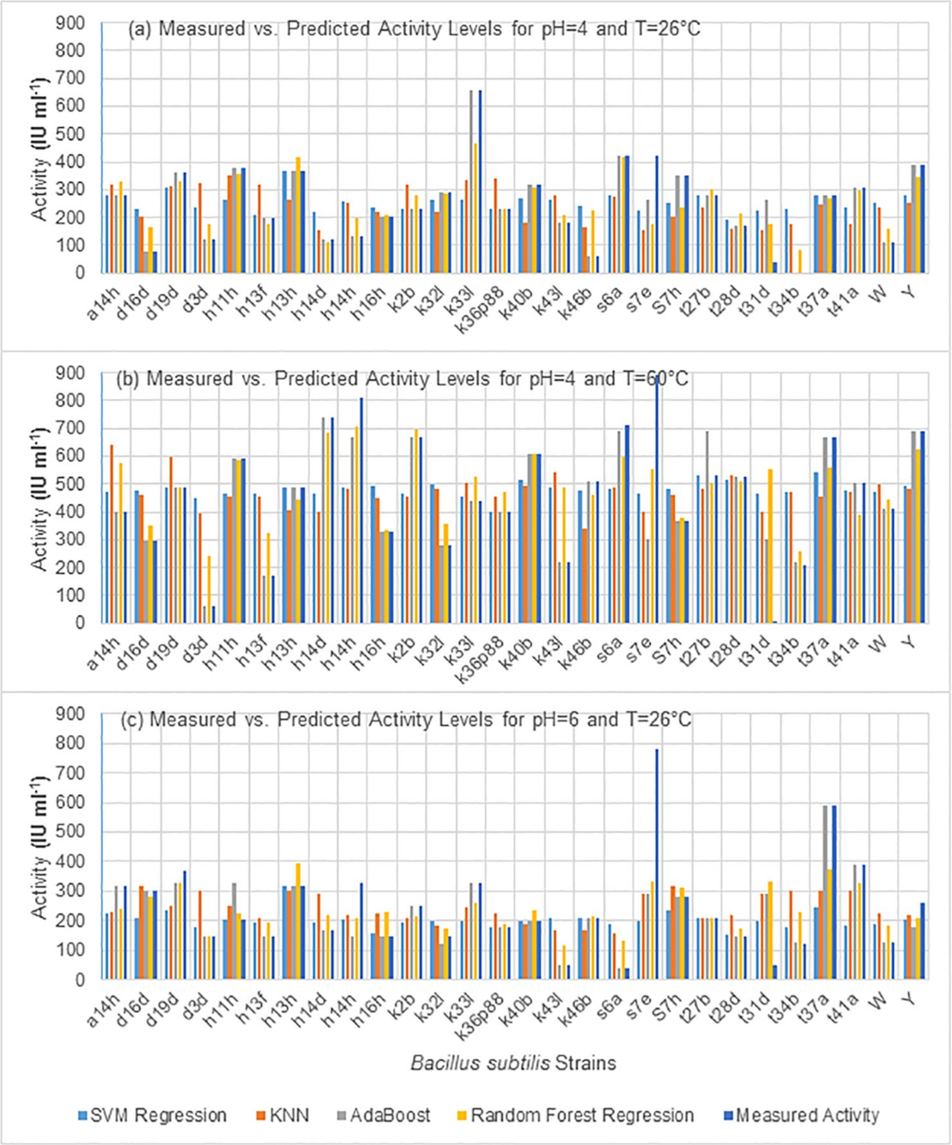

Fig 2. The activity of xylanase enzymes purified from different *Bacillus subtilis* strains vs. predicted activities by four computational models. The activities were determined in three different pH/temperature conditions. (a) pH = 4, T = 26˚C (b) pH = 4,T = 60˚C (c) pH = 6,T = 26˚C.

https://doi.org/10.1371/journal.pone.0205796.g002

dozens of enzymes from the same family, with known amino acid sequence and precise specific activity values in the same pH and temperature to learn our regression model.

Despite many empirical studies which have been done to measure the activity of a variety of enzymes, due to the lack of enough proper training data for specific temperature/pH condition, very little has been done to build statistical learner models for activity prediction from sequence.

This research work is one of the primary steps to cover this deficiency. Due to the fact that screening the bacterial halos is a primary step for selecting proper strains, a method was proposed that can classify the magnitude of the diameter of bacterial halo zone by exploiting PseAAC features. In the xylanase selective medium, the halo diameter is highly correlated with

Table 3. Performance measures resulted from four different regression models. The models were validated by stratified 10-fold cross validation and jackknife methods. The results are related to three different pH/Temperature conditions. (a) pH = 4,T = 26˚C (b) pH = 4,T = 60˚C (c) pH = 6,T = 26˚C.

			Regression Models							
			SVM		***KNN (K = 5)***		***AdaBoost***		***Random Forest Regression***	
			10-fold	***Jackknife***	***10-fold***	***Jackknife***	***10-fold***	***Jackknife***	***10-fold***	***Jackknife***
Assay Conditions	pH = 4 T = 26˚C	MSE	24529.116	22879.562	25600.679	26193.536	35004.789	39039.153	34166.515	29285.824
		RMSE	156.618	151.260	160.002	161.844	187.096	197.583	184.842	171.131
		MAE	126.693	122.028	135.536	135.893	147.176	162.334	139.415	134.655
		R2	-0.21	-0.128	-0.263	-0.292	-0.726	-0.925	-0.685	-0.444
	pH = 4 T = 60˚C	MSE	55753.178	57096.944	72585.964	73541.250	123725.753	114543.750	78659.703	77522.968
		RMSE	236.121	238.950	269.418	271.185	351.747	338.443	280.463	278.429
		MAE	197.666	199.404	227.179	231.821	284.171	254.821	222.875	226.616
		R2	-0.192	-0.221	-0.552	-0.572	-1.646	-1.449	-0.682	-0.658
	pH = 6 T = 26˚C	MSE	31188.785	29521.580	32759.964	32268.250	62414.683	59070.759	44906.883	44090.080
		RMSE	176.603	171.818	180.997	179.634	249.829	243.045	211.912	209.976
		MAE	130.15	126.440	134.179	132.321	182.262	160.089	149.517	159.520
		R2	-0.282	-0.213	-0.347	-0.326	-1.565	-1.428	-0.846	-0.812

https://doi.org/10.1371/journal.pone.0205796.t003

the activity of the expressed xylanase from the corresponding strain. However, it is clear that the halo diameter is not the only function of the xylanase activity, but, many factors play a role in its formation. Therefore, the exact prediction of halo diameter only from xylanase sequence is impossible. Nevertheless, we showed that correct classification of halo diameter from the enzyme sequence in one of H, M, and L classes is logical and feasible. Therefore, we developed two learning models which help to obtain a relatively accurate estimation of activity for new xylanases and bacterial halos diameter for new strains without the need for new experiments. Finding a reliable in-silico prediction model for enzyme function and activity, may circumvent costly and time consuming experimental screening. We showed that the problem of enzyme activity prediction solely from its primary structure could be partially solved by regression machines. Adequate training data makes the regression results more reliable and informative. However, the accuracy and precision of predictors remain as serious concerns. The main reason is that choosing a model because of its performance based on limited training data, does not guarantee the correct prediction of future observations, also known as the generalization power of predictor. Cross-validation is a technique for evaluating predictive models and assesses how the performance of a learning model will be generalized to independent and unseen datasets. Our proposed models were validated using stratified cross-validation and the jackknife techniques.

Although PseAAC features have been used in many prediction tasks in computational biology, to the best of our knowledge, its usage for determining the activity of enzymes in specific conditions has not been reported yet. Further efforts are required to develop similar computational models for enzyme activity prediction based on the other bio-physicochemical and evolutionary features that can be extracted from the amino acid sequence of enzymes. The features obtained from PSSM (Position Specific Scoring Matrix), hydrophobicity, polarity, polarizability, and many others are among such sources of information about enzyme activity.

In this work we used a feature vector with 34 elements. Using other information sources such as the above mentioned features, can heavily increase the feature vector dimension. In machine learning tasks, high dimension feature will maybe result in three problems: one is over-fitting which results in low generalization ability of prediction model; another is information redundancy or noise which results in bad prediction accuracy; the other is dimension disaster which results in a handicap for the computation. Using feature selection techniques to optimize feature set can not only economize the time for computation, but also build robust prediction model. In fact, many techniques such as principal component analysis (PCA) [53], minimal-redundancy-maximal-relevance (mRMR) [54], analysis of variance (ANOVA) [55], F-score algorithm [40], binomial distribution [56] have been proposed and used in sequence analysis and prediction. Thus, feature selection in the future works hopefully can improve prediction results.

As main achievement, the proposed methodology can be used for any family of enzymes, with exploiting any kind of regression machine and any sequence based feature vectors other than those discussed in this work. The only limitation is the availability of sufficient training data for specified temperature and pH condition.

No single general computational approach alone is likely to be a perfect solution for the problem of predicting the activity of homologous enzymes from different families [50]. However, it is possible and plausible for a specific family of enzymes to determine the activity of some members based on the determined activities of the others. In the current state, the lack of computational tools with the capability of enzyme activity prediction is tangible in both scientific studies and industrial applications. Considering the diversity of enzyme families and large number of members in each family, it seems very difficult to design a general purpose machine that can accurately predict enzyme activity in different pH and temperature conditions only from sequence based data. It is more practical to design and implement a special purpose predictor machine for each family of enzymes. These machines can be trained based on experimental activity measurements and evaluated with proper testing datasets. One of the main applications of predictive models similar to those introduced in this work is to select new suitable candidate enzymes with superior activities from huge metagenome data. It is almost impossible to select good targets without automated activity prediction tools. Since user-friendly and publicly accessible web-servers represent the future direction for developing practically more useful models [38,81–86], more efforts will be made in the future work to provide a web-server for the method presented in this paper.

Supporting information

S1 Table. The PseAAC feature values. Features extracted from 41 studied xylanase amino acid sequences. $\lambda = 7$ was used for generating features.
(XLSX)

S2 Table. Studied xylanase sequences. The amino acid sequences of 41 different xylanase enzymes and their GenBank Accession No. are provided.
(XLSX)

Acknowledgments

The authors would like to thank Dr. Reza Ghaffari for his comprehensive support of the project and sharing his wisdom with us during the course of this research.

Author Contributions

Data curation: Maryam Mousivand, Parinaz Moradi Dezfouli, Maryam Hashemi, Kaveh Kavousi.

Formal analysis: Shohreh Ariaeenejad.

Investigation: Maryam Hashemi, Ghasem Hosseini Salekdeh.

Methodology: Shohreh Ariaeenejad, Maryam Mousivand, Parinaz Moradi Dezfouli, Kaveh Kavousi.

Project administration: Shohreh Ariaeenejad.

Resources: Maryam Hashemi.

Software: Shohreh Ariaeenejad.

Supervision: Ghasem Hosseini Salekdeh.

Validation: Maryam Hashemi, Ghasem Hosseini Salekdeh.

Visualization: Maryam Mousivand.

Writing – original draft: Shohreh Ariaeenejad.

Writing – review & editing: Ghasem Hosseini Salekdeh.

References

1. Henrissat B, Bairoch A (1993) New families in the classification of glycosyl hydrolases based on amino acid sequence similarities. Biochemical Journal 293: 781–788. PMID: 8352747
2. Henrissat B, Bairoch A (1996) Updating the sequence-based classification of glycosyl hydrolases. Biochemical Journal 316: 695. PMID: 8687420
3. Jeffries TW (1996) Biochemistry and genetics of microbial xylanases. Current opinion in Biotechnology 7: 337–342. PMID: 8785441
4. Ohmiya K, Sakka K, Karita S, Kimura T (1997) Structure of cellulases and their applications. Biotechnology and Genetic Engineering Reviews 14: 365–414. PMID: 9188160
5. Viikari L, Kantelinen A, Sundquist J, Linko M (1994) Xylanases in bleaching: from an idea to the industry. FEMS Microbiology Reviews 13: 335–350.
6. Gilbert HJ, Hazlewood GP (1993) Bacterial cellulases and xylanases. Microbiology 139: 187–194.
7. Coughlan M, Hazlewood GP (1993) beta-1, 4-D-xylan-degrading enzyme systems: biochemistry, molecular biology and applications. Biotechnology and Applied Biochemistry 17: 259–289. PMID: 8338637
8. Chen X, Whitmire D, Bowen JP (1996) Xylanase homology modeling using the inverse protein folding approach. Protein science 5: 705–708. https://doi.org/10.1002/pro.5560050415 PMID: 8845760
9. Sá-Pereira P, Paveia H, Costa-Ferreira M, Aires-Barros MR (2003) A new look at xylanases. Molecular biotechnology 24: 257–281. PMID: 12777693
10. Elegir G, Szakács G, Jeffries TW (1994) Purification, characterization, and substrate specificities of multiple xylanases from Streptomyces sp. strain B-12-2. Applied and Environmental Microbiology 60: 2609–2615. PMID: 16349337
11. Thomas L, Ushasree MV, Pandey A (2014) An alkali-thermostable xylanase from Bacillus pumilus functionally expressed in Kluyveromyces lactis and evaluation of its deinking efficiency. Bioresource technology 165: 309–313. https://doi.org/10.1016/j.biortech.2014.03.037 PMID: 24709528
12. Cintra LC, Fernandes AG, de Oliveira ICM, Siqueira SJL, Costa IGO, Colussi F, et al. (2017) Characterization of a recombinant xylose tolerant-xylosidase from Humicola grisea var. thermoidea and its use in sugarcane bagasse hydrolysis.
13. Zheng H, Liu Y, Sun M, Han Y, Wang J, Lu F, et al. (2014) Improvement of alkali stability and thermostability of Paenibacillus campinasensis Family-11 xylanase by directed evolution and site-directed mutagenesis. Journal of industrial microbiology & biotechnology 41: 153–162.
14. Basu M, Kumar V, Shukla P (2018) Recombinant approaches for microbial xylanases: Recent advances and perspectives. Current Protein and Peptide Science 19: 87–99. PMID: 27875966
15. Song J, Tan H, Mahmood K, Law RH, Buckle AM, Webb GI, et al. (2009) Prodepth: predict residue depth by support vector regression approach from protein sequences only. PloS one 4: e7072. https://doi.org/10.1371/journal.pone.0007072 PMID: 19759917
16. Hediger MR, De Vico L, Svendsen A, Besenmatter W, Jensen JH (2012) A computational methodology to screen activities of enzyme variants. PloS one 7: e49849. https://doi.org/10.1371/journal.pone.0049849 PMID: 23284627

17. Sá-Pereira P, Mesquita A, Duarte JC, Barros MRA, Costa-Ferreira M (2002) Rapid production of thermostable cellulase-free xylanase by a strain of Bacillus subtilis and its properties. Enzyme and Microbial Technology 30: 924–933.

18. Huang J, Wang G, Xiao L (2006) Cloning, sequencing and expression of the xylanase gene from a Bacillus subtilis strain B10 in Escherichia coli. Bioresource Technology 97: 802–808. https://doi.org/10.1016/j.biortech.2005.04.011 PMID: 15951169

19. Nakamura S, Wakabayashi K, Nakai R, Aono R, Horikoshi K (1993) Purification and some properties of an alkaline xylanase from alkaliphilic Bacillus sp. strain 41M-1. Applied and Environmental Microbiology 59: 2311–2316. PMID: 8292206

20. Bernier R, Driguez H, Desrochers M (1983) Molecular cloning of a Bacillus subtilis xylanase gene in Escherichia coli. Gene 26: 59–65. PMID: 6423449

21. Bernier R, Desrochers M, Jurasek L, Paice MG (1983) Isolation and characterization of a xylanase from Bacillus subtilis. Applied and environmental microbiology 46: 511–514. PMID: 16346375

22. Jalal A, Rashid N, Rasool N, Akhtar M (2009) Gene cloning and characterization of a xylanase from a newly isolated Bacillus subtilis strain R5. Journal of bioscience and bioengineering 107: 360–365. https://doi.org/10.1016/j.jbiosc.2008.12.005 PMID: 19332293

23. Jiang Z, Wei Y, Li D, Li L, Chai P, Kusakabe I, et al. (2006) High-level production, purification and characterization of a thermostable β-mannanase from the newly isolated Bacillus subtilis WY34. Carbohydrate Polymers 66: 88–96.

24. Chou K-C (2005) Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21: 10–19. https://doi.org/10.1093/bioinformatics/bth466 PMID: 15308540

25. Chou KC (2001) Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins: Structure, Function, and Bioinformatics 43: 246–255.

26. Lai H-Y, Chen X-X, Chen W, Tang H, Lin H (2017) Sequence-based predictive modeling to identify cancerlectins. Oncotarget 8: 28169. https://doi.org/10.18632/oncotarget.15963 PMID: 28423655

27. Yang H, Tang H, Chen X-X, Zhang C-J, Zhu P-P, Ding H, et al. (2016) Identification of secretory proteins in mycobacterium tuberculosis using pseudo amino acid composition. BioMed research international.

28. Tang H, Chen W, Lin H (2016) Identification of immunoglobulins using Chou's pseudo amino acid composition with feature selection technique. Molecular BioSystems 12: 1269–1275. https://doi.org/10.1039/c5mb00883b PMID: 26883492

29. Tang H, Su Z-D, Wei H-H, Chen W, Lin H (2016) Prediction of cell-penetrating peptides with feature selection techniques. Biochemical and biophysical research communications 477: 150–154. https://doi.org/10.1016/j.bbrc.2016.06.035 PMID: 27291150

30. Chen X-X, Tang H, Li W-C, Wu H, Chen W, Ding H, et al. (2016) Identification of bacterial cell wall lyases via pseudo amino acid composition. BioMed research international 2016.

31. Lin H, Deng E-Z, Ding H, Chen W, Chou K-C (2014) iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. Nucleic acids research 42: 12961–12972. https://doi.org/10.1093/nar/gku1019 PMID: 25361964

32. Guo S-H, Deng E-Z, Xu L-Q, Ding H, Lin H, Chen W, et al. (2014) iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition. Bioinformatics 30: 1522–1529. https://doi.org/10.1093/bioinformatics/btu083 PMID: 24504871

33. Liu B, Long R, Chou K-C (2016) iDHS-EL: identifying DNase I hypersensitive sites by fusing three different modes of pseudo nucleotide composition into an ensemble learning framework. Bioinformatics 32: 2411–2418. https://doi.org/10.1093/bioinformatics/btw186 PMID: 27153623

34. Liu B, Fang L, Long R, Lan X, Chou K-C (2015) iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition. Bioinformatics 32: 362–369. https://doi.org/10.1093/bioinformatics/btv604 PMID: 26476782

35. Chen W, Feng P-M, Deng E-Z, Lin H, Chou K-C (2014) iTIS-PseTNC: a sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition. Analytical biochemistry 462: 76–83. https://doi.org/10.1016/j.ab.2014.06.022 PMID: 25016190

36. Chen W, Feng P-M, Lin H, Chou K-C (2013) iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. Nucleic acids research 41: e68–e68. https://doi.org/10.1093/nar/gks1450 PMID: 23303794

37. Chen W, Feng P-M, Lin H, Chou K-C (2014) iSS-PseDNC: identifying splicing sites using pseudo dinucleotide composition. BioMed research international.

38. Chen W, Tang H, Ye J, Lin H, Chou K-C (2016) iRNA-PseU: Identifying RNA pseudouridine sites. Molecular Therapy—Nucleic Acids 5: e332. PMID: 28427142

39. Zhang C-J, Tang H, Li W-C, Lin H, Chen W, Chou K, et al. (2016) iOri-Human: identify human origin of replication by incorporating dinucleotide physicochemical properties into pseudo nucleotide composition. Oncotarget 7: 69783–69793. https://doi.org/10.18632/oncotarget.11975 PMID: 27626500

40. Lin H, Liang Z-Y, Tang H, Chen W (2017) Identifying sigma70 promoters with novel pseudo nucleotide composition. IEEE/ACM transactions on computational biology and bioinformatics.

41. Chen W, Lei T-Y, Jin D-C, Lin H, Chou K-C (2014) PseKNC: a flexible web server for generating pseudo K-tuple nucleotide composition. Analytical biochemistry 456: 53–60. https://doi.org/10.1016/j.ab.2014.04.001 PMID: 24732113

42. Chen W, Zhang X, Brooker J, Lin H, Zhang L, Chou K, et al. (2014) PseKNC-General: a cross-platform package for generating various modes of pseudo nucleotide compositions. Bioinformatics 31: 119–120. https://doi.org/10.1093/bioinformatics/btu602 PMID: 25231908

43. Liu B, Liu F, Wang X, Chen J, Fang L, Chou K, et al. (2015) Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. Nucleic acids research 43: W65–W71. https://doi.org/10.1093/nar/gkv458 PMID: 25958395

44. Chou K-C, Zhang C-T, Maggiora GM (1997) Disposition of amphiphilic helices in heteropolar environments. Proteins Structure Function and Genetics 28: 99–108.

45. Shen H-B, Chou K-C (2006) Ensemble classifier for protein fold pattern recognition. Bioinformatics 22: 1717–1722. https://doi.org/10.1093/bioinformatics/btl170 PMID: 16672258

46. Shen H-B, Chou K-C (2008) PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition. Analytical biochemistry 373: 386–388. https://doi.org/10.1016/j.ab.2007.10.012 PMID: 17976365

47. Feng P, Ding H, Yang H, Chen W, Lin H, Chou K, et al. (2017) iRNA-PseColl: Identifying the Occurrence Sites of Different RNA Modifications by Incorporating Collective Effects of Nucleotides into PseKNC. Molecular Therapy-Nucleic Acids 7: 155–163. https://doi.org/10.1016/j.omtn.2017.03.006 PMID: 28624191

48. Feng P-M, Ding H, Chen W, Lin H (2013) Naive Bayes classifier with feature selection to identify phage virion proteins. Computational and mathematical methods in medicine 2013.

49. Zuo Y-C, Su W-X, Zhang S-H, Wang S-S, Wu C-Y, Yang L, et al. (2015) Discrimination of membrane transporter protein types using K-nearest neighbor method derived from the similarity distance of total diversity measure. Molecular bioSystems 11: 950–957. https://doi.org/10.1039/c4mb00681j PMID: 25607774

50. Jacobson MP, Kalyanaraman C, Zhao S, Tian B (2014) Leveraging structure for enzyme function prediction: methods, opportunities, and challenges. Trends in biochemical sciences 39: 363–371. PMID: 24998033

51. Kavousi K, Moshiri B, Sadeghi M, Araabi BN, Moosavi-Movahedi AA (2011) A protein fold classifier formed by fusing different modes of pseudo amino acid composition via PSSM. Computational biology and chemistry 35: 1–9. https://doi.org/10.1016/j.compbiolchem.2010.12.001 PMID: 21216672

52. Kavousi K, Sadeghi M, Moshiri B, Araabi BN, Moosavi-Movahedi AA (2012) Evidence theoretic protein fold classification based on the concept of hyperfold. Mathematical Biosciences 240: 148–160. https://doi.org/10.1016/j.mbs.2012.07.001 PMID: 22824139

53. Li Z, Wang J, Zhang S, Zhang Q, Wu W (2017) A new hybrid coding for protein secondary structure prediction based on primary structure similarity. Gene 618: 8–13. https://doi.org/10.1016/j.gene.2017.03.011 PMID: 28322997

54. Huo H, Li T, Wang S, Lv Y, Zuo Y, Yang L, et al. (2017) Prediction of presynaptic and postsynaptic neurotoxins by combining various Chou's pseudo components. Scientific Reports. 7.

55. Zhao Y-W, Su Z-D, Yang W, Lin H, Chen W, Tang H, et al. (2017) IonchanPred 2.0: A Tool to Predict Ion Channels and Their Types. International Journal of Molecular Sciences 18: 1838.

56. Zhu P-P, Li W-C, Zhong Z-J, Deng E-Z, Ding H,Chen W, et al. (2015) Predicting the subcellular localization of mycobacterial proteins by incorporating the optimal tripeptides into the general form of pseudo amino acid composition. Molecular BioSystems 11: 558–563. https://doi.org/10.1039/c4mb00645c PMID: 25437899

57. Chen W, Feng P, Ding H, Lin H, Chou K-C (2015) iRNA-methyl: identifying N 6-methyladenosine sites using pseudo nucleotide composition. Analytical biochemistry 490: 26–33. https://doi.org/10.1016/j.ab.2015.08.021 PMID: 26314792

58. Tang H, Zou P, Zhang C, Chen R, Chen W, Lin H, et al. (2016) Identification of apolipoprotein using feature selection technique. Scientific reports. 6.

59. Liu B, Wang S, Long R, Chou K-C (2016) iRSpot-EL: identify recombination spots with an ensemble learning approach. Bioinformatics 33: 35–41. https://doi.org/10.1093/bioinformatics/btw539 PMID: 27531102

60. Liu B, Yang F, Chou K-C (2017) 2L-piRNA: A Two-Layer Ensemble Classifier for Identifying Piwi-Interacting RNAs and Their Function. Molecular Therapy-Nucleic Acids 7: 267–277. https://doi.org/10.1016/j.omtn.2017.04.008 PMID: 28624202

61. Khandeparker R, Verma P, Deobagkar D (2011) A novel halotolerant xylanase from marine isolate Bacillus subtilis cho40: gene cloning and sequencing. New biotechnology 28: 814–821. https://doi.org/10.1016/j.nbt.2011.08.001 PMID: 21890005

62. Tariq R, Ansari I, Qadir F, Ahmed A, Shariq M, Zafar U, et al. (2018) Optimization of Endoglucanase Production From Thermophilic Strain of Bacillus Licheniformis RT-17 and ITS Application for Saccharification of Sugarcane Bagasse. Pakistan Journal of Botany 50: 807–816.

63. Kumar V, Dangi AK, Shukla P (2018) Engineering thermostable microbial xylanases toward its industrial applications. Molecular biotechnology: 1–10.

64. Shukla P (2018) 'Futuristic Protein Engineering: Developments and Avenues'. Current Protein and Peptide Science 19: 3–4. PMID: 29197322

65. Miller GL (1959) Use oi Dinitrosalicylic Acid Reagent tor Determination oi Reducing Sugar. Analytical chemistry 31: 426–428.

66. Li W-C, Deng E-Z, Ding H, Chen W, Lin H (2015) iORI-PseKNC: a predictor for identifying origin of replication with pseudo k-tuple nucleotide composition. Chemometrics and Intelligent Laboratory Systems 141: 100–106.

67. Liu B, Wu H, Zhang D, Wang X, Chou K-C (2017) Pse-Analysis: a python package for DNA/RNA and protein/peptide sequence analysis based on pseudo components and kernel methods. Oncotarget 8: 13338. https://doi.org/10.18632/oncotarget.14524 PMID: 28076851

68. Chou K-C (2000) Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. Biochemical and biophysical research communications 278: 477–483. https://doi.org/10.1006/bbrc.2000.3815 PMID: 11097861

69. Chen W, Yang H, Feng P, Ding H, Lin H (2017) iDNA4mC: identifying DNA N4-methylcytosine sites based on nucleotide chemical properties. Bioinformatics.

70. Demšar J, Curk T, Erjavec A, Gorup Č, Hočevar T,Milutinovic M, et al. (2013) Orange: data mining toolbox in Python. The Journal of Machine Learning Research 14: 2349–2353.

71. Tang H, Yang Y, Zhang C, Chen R, Huang P, Duan C, et al. (2017) Predicting Presynaptic and Postsynaptic Neurotoxins by Developing Feature Selection Technique. BioMed Research International 2017.

72. Ding H, Li D (2015) Identification of mitochondrial proteins of malaria parasite using analysis of variance. Amino acids 47: 329–333. https://doi.org/10.1007/s00726-014-1862-4 PMID: 25385313

73. Zhao Y-W, Lai H-Y, Tang H, Chen W, Lin H (2016) Prediction of phosphothreonine sites in human proteins by fusing different features. Scientific reports 6.

74. Chou K-C, Zhang C-T (1995) Prediction of protein structural classes. Critical reviews in biochemistry and molecular biology 30: 275–349. https://doi.org/10.3109/10409239509083488 PMID: 7587280

75. Ding H, Feng P-M, Chen W, Lin H (2014) Identification of bacteriophage virion proteins by the ANOVA feature selection and analysis. Molecular BioSystems 10: 2229–2235. https://doi.org/10.1039/c4mb00316k PMID: 24931825

76. Lin H, Chen W (2011) Prediction of thermophilic proteins using feature selection technique. Journal of microbiological methods 84: 67–70. https://doi.org/10.1016/j.mimet.2010.10.013 PMID: 21044646

77. Yuan L-F, Ding C, Guo S-H, Ding H, Chen W, Lin H, et al. (2013) Prediction of the types of ion channel-targeted conotoxins based on radial basis function network. Toxicology in Vitro 27: 852–856. https://doi.org/10.1016/j.tiv.2012.12.024 PMID: 23280100

78. Lin H, Ding H, Guo F-B, Zhang A-Y, Huang J (2008) Predicting subcellular localization of mycobacterial proteins by using Chou's pseudo amino acid composition. Protein and peptide letters 15: 739–744. PMID: 18782071

79. Ding H, Luo L, Lin H (2009) Prediction of cell wall lytic enzymes using Chou's amphiphilic pseudo amino acid composition. Protein and peptide letters 16: 351–355. PMID: 19356130

80. Cuesta SM, Rahman SA, Furnham N, Thornton JM (2015) The classification and evolution of enzyme function. Biophysical journal 109: 1082–1086. https://doi.org/10.1016/j.bpj.2015.04.020 PMID: 25986631

81. Liang Z-Y, Lai H-Y, Yang H, Zhang C-J, Yang H, Wei H, et al. (2017) Pro54DB: a database for experimentally verified sigma-54 promoters. Bioinformatics 33: 467–469. https://doi.org/10.1093/bioinformatics/btw630 PMID: 28171531

82. Zhang T, Tan P, Wang L, Jin N, Li Y, Zhang L, et al. (2017) RNALocate: a resource for RNA subcellular localizations. Nucleic acids research 45: D135–D138. https://doi.org/10.1093/nar/gkw728 PMID: 27543076

83. Chen W, Tang H, Lin H (2017) MethyRNA: a web server for identification of N6-methyladenosine sites. Journal of Biomolecular Structure and Dynamics 35: 683–687. https://doi.org/10.1080/07391102.2016.1157761 PMID: 26912125

84. Ding H, Yang W, Tang H, Feng P-M, Huang J, Lin H, et al. (2016) PHYPred: a tool for identifying bacteriophage enzymes and hydrolases. Virologica Sinica 31: 350. https://doi.org/10.1007/s12250-016-3740-6 PMID: 27151186

85. Ding H, Deng E-Z, Yuan L-F, Liu L, Lin H,Wei C, et al. (2014) iCTX-Type: a sequence-based predictor for identifying the types of conotoxins in targeting ion channels. BioMed research international.

86. Chen W, Lin H, Feng P-M, Ding C, Zuo Y-C, Chou K, et al. (2012) iNuc-PhysChem: a sequence-based predictor for identifying nucleosomes via physicochemical properties. PloS one 7: e47843. https://doi.org/10.1371/journal.pone.0047843 PMID: 23144709