BioTechnologia vol. 95(2) • pp. 161-173 • 2014
Journal of Biotechnology, Computational Biology and Bionanotechnology RESEARCH PAPER

Determining the structural amino acid attributes which are important in both protein thermostability and alkalophilicity: a case study on xylanase

AZAR DELAVARI[1], SAJJAD ZARE[2], MOHAMMAD REZA GHAEMI[3], RAFIEH KASHFI[2], MAHDI EBRAHIMI[4], AHMAD TAHMASEBI[2], MANSOUR EBRAHIMI[5], ESMAEIL EBRAHIMIE[2, 6]*

[1] Institute of Biotechnology, College of Agriculture, Shiraz University, Iran

[2] Department of Crop Production & Plant Breeding, College of Agriculture, Shiraz University, Iran

[3] Department of Health Behavior, Roswell Park Cancer Institute, Buffalo, NY 14263, USA

[4] Alumnus from Department of Informatics, Saarland University, Saarbrucken, Germany

[5] Bioinformatics Research Group, Department of Biological Sciences, Qom University, Iran

[6] Discipline of Genetics, School of Molecular and Biomedical Science, The University of Adelaide, Adelaide, Australia

*Corresponding author: esmaeil.ebrahimie@adelaide.edu.au

Abstract

Xylanases are used in the recycling of biomass and have other industrial applications including pulp bleaching. These enzymes are also applied in the baking industry and for the manufacture of animal feed. Such technologies as, for example, pulp bleaching entail high temperatures and high pHs. As a result, there is great demand from industry for thermostable and halostable forms of xylanase. Due to the relatively high variation in the thermo- and halo-stability of xylanases, feature selection was employed as a model to discover the important attributes of their amino acid sequences affecting the thermo- and halo-stability of the enzyme. A data set containing the amino acid sequences of xylanases with different thermo- and halostabilities was collected. Seventy-four amino acid attributes were obtained for each enzyme sequence. After running a feature selection algorithm for each of the thermo- and halostablity variables, features were classified as either important, unimportant or marginal. The results showed a significant correlation between structural amino acid attitudes and stability in harsh temperatures or alkaline conditions. Features such as lysine, glutamic acid, and positively/negatively charged residues showed a positive correlation with both the thermostability and alkalophilicity attributes of the protein. For the first time, we found attributes which were important for both stability at high temperatures as well as in alkaline conditions by mining sequence-derived amino acid attributes using data mining.

Key words: amino acid, bioinformatics, data mining, feature selection

Introduction

Xylan is the major hemicellulose component of wood which can constitute up to 35% of the plant's total dry weight (Decelle et al., 2004). Due to its structural complexity (Cheng et al., 2008), hydrolyzing xylan into simple sugars requires the synergistic actions of a series of enzymes. Regarding the specificity of xylan degradation, endo-β-1,4-xylanase (β-1,4-D-xylan xylanohydrolase; EC 3.2.1.8) has been named a "true xylanase" within the various xylanolytic enzymes (Collins et al., 2006). Xylanase (EC3.2.1.8) is an important enzyme in the recycling of biomass with a wide usage in industrial biotechnology in, e.g., pulp bleaching, the baking industry, and the manufacture of animal feed. Xylanases used in the paper industry are eco-friendly, because they reduce the amount of toxic chlorine in the pulp pre-bleaching process (Natesh et al., 1999).

Since pulp bleaching entails high temperatures and high pHs, the optimum xylanase used has to be thermophilic, thermostable, alkalophilic, and stable in alkaline environments (Shoham et al., 1992). To date, different approaches have been adopted to improve the thermo- and halostability of xylanases. Methods for inducing random mutagenesis (error-prone PCR and chemical mutagenesis) have been applied to improve the properties of xylanase. The thermostability of xylanase has been incre-

ased through chemical mutagenesis (Arase et al., 1993). Stabilizing amino acid substitutions have provided an insight into the conservation of valine at position 334 and polar residues in the vicinity of residue 348 in GH10 xylanases (Xie et al., 2006). A Directed mutation has been used to produce an alkalophilic variant from a *Neocallimastix patriciarum* xylanase (Chen et al., 2001). A site directed mutagenesis has been shown to be effective for compiling the eight amino acid substitutions into a series of composite mutant xylanases (Chen et al., 2001). However, only xyn-CDBFV (with seven amino acid mutations) has proved to be more alkalophilic.

Even though protein engineering is the most powerful tool for redesigning proteins, the reports on xylanases suggest that the protein properties of xylanases have not been fully exploited to provide the required information for protein engineering (Kulkarni et al., 2006). Feature selection is an essential tool with many applications, including bioinformatics where we encounter high-dimensional data (Bakhtiarizadeh et al., 2014; Ebrahimi et al., 2014; Zinati et al., 2014). During the last decade, the motivation for applying feature selection (attribute weighting) techniques in bioinformatics has shifted from being an illustrative example to becoming a real prerequisite for model building (Krämer et al., 2009). The main goal of feature selection is to extract the most relevant information from the primary feature set. In previous studies, feature selection has demonstrated its appropriacy as a highly reliable data mining technique in discovering the main features for discriminating mesostable from thermostable enzymes (Lakizadeh et al., 2011), finding the protein attributes contributing to halostability (Ebrahimi and Ebrahimie, 2010; Ebrahimie et al., 2011), defining the most important attitudes affecting physiological traits (Shekoofa et al., 2011), grain yield in plants (Bijanzadeh et al., 2010), and even in discriminating malignant from benign breast cancer (Ebrahimi et al., 2010). Moreover, feature selection has been efficiently employed to increase the accuracy of different prediction models and to remove the redundancy in data from protein thermostability (Ebrahimi et al., 2010) and P1B-ATPase heavy metal transporters (Ashrafi et al., 2011). It is worth mentioning that feature selection is not only helpful for reducing the computational burden of analyzing data, but it also markedly reduces the time spent in the laboratory examining hundreds of features.

However, to date, there has been no comparative analysis of halo- and thermostability traits within a given class of enzyme sequences using the feature selection method. In this study, feature selection has been used to identify the important attributes which affect the thermostability and alkalophilicity of xylanase sequences. Statistical analyses helped demonstrate negative or positive relationships between selected features based on pH and temperature tolerance. A major feature which has made this work possible is that each sequence has specific characteristics, which are stable at defined temperatures and pHs. These characteristics can be identified using attribute weighting on xylanases composed of amino acids containing different thermo- and halostabilities.

Materials and methods

Data Collection and preparation

To obtain xylanase stability characteristics, a dataset containing 90 xylanase protein sequences, dissimilar in thermo- and halostabilities, was constructed from the Swiss-Prot and TrEMBL databases (http:au.expasy. org). Of the 90 xylanases from the available literature used in our study, the temperatures and pHs of only 81 xylanases had been reported previously. Ten amylases were added to enrich the data set. Information about these enzymes was extracted from the Swiss-Prot and TrEMBL databases and from publications (See supplementary 1.xlsx for details).

Computational calculation of amino acid features and feature ranking

Each amino acid sequence has specific and individual features which are unique to the particular xylanase. Features such as frequency or the Count of amino acids can reveal important information about a novel protein. Seventy-four features for each enzyme were extracted using the CLC bio tool (http:www.clcbio.com). These features included: length, weight (KDa), isoelectric point, aliphatic index, N-terminal amino acid, half-life in: mammals (hours), yeast (> hours) and *E. coli* (> hours), non-reduced cysteine extinction coefficient, non-reduced cysteine absorption, reduced cysteine extinction coefficient, reduced cysteine absorption, sulfur count, carbon count, nitrogen count, oxygen count, hydrogen count, sulfur frequency, carbon frequency, nitrogen frequency, oxygen frequency, hydrogen frequency, hydrophobic re-

sidue count, hydrophilic residue count, other residue count, hydrophobic residue frequency, hydrophilic residue frequency, other residue frequency, negatively charged residue count, negatively charged residue frequency, positively charged residue frequency, counts of: Ala(alanine), Cys(cysteine), Asp(aspartic acid), Glu(glutamic acid), Phe(phenylalanine), Gly(glycine), His(histidine), Ile(isoleucine), Lys(lysine), Leu(leucine), Met(methionine), Asn(asparagine), Pro(proline), Gln(glutamine) Arg(arginine), Ser(serine), Thr(threonine), Val(valine), Trp(tryptophan), Tyr(tyrosine), frequencies of: Ala, Cys, Asp, Glu, Phe, Gly, His, Ile, Lys, Leu, Met, Asn(asparagine), Pro, Gln, Arg, Ser, Thr, Val, Trp, Tyr.

Description of the features

The molecular weight of a protein is simply calculated from the sum of the atomic masses of all the atoms in the molecule. The weight of a protein is usually represented in Daltons (Da).

The isoelectric point (pI) of a protein is the pH at which proteins have no net charge. The pI is calculated from the pKa values of 20 different amino acids. At a pH below pI, proteins carry a positive charge, whereas if the pH is above pI the proteins carry a negative charge.

The Aliphatic index of a protein is a measure of the relative volume occupied by the aliphatic side chain of the following amino acids: Ala, Val, Leu and Ile. An increase in the aliphatic index increases the thermostability of globular proteins. The index is calculated *via* the following formula:

$$\text{Aliphatic index} = X(\text{Ala}) + a \times X(\text{Val}) + b \times X(\text{Leu}) + b \times (X)\text{Ile}$$

The constants a and b are the relative volume of valine (a = 2.9) and leucine/isoleucine (b = 3.9) side chains compared to the side chain of alanine.

The estimated half-life of a protein is the time it takes for the protein pool of that particular protein to be reduced by half. The half-life of a protein is highly dependent on the presence of the N-terminal amino acid.

The extinction coefficient is the measure indicating how much light is absorbed by a protein at a particular wave length. The extinction coefficient is calculated from the absorbance of cysteine, tyrosine and tryptophan using the following equation:

$$\text{Ext(Protein)} = \text{count(Cys)} \times \text{Ext(Cys)} + \text{count(Tyr)} \times \text{Ext(Tyr)} + \text{count(Trp)} \times \text{Ext(Trp)}.$$

At 280 nm, the extinction coefficients are: Cys =120, Tyr =1280 and Trp = 5690. On the basis of the extinction coefficient, the absorbance (optical density) can be calculated using the following formula:

$$\text{Absorbance(Protein)} = \text{Ext(Protein)} / \text{Molecular weight}.$$

Atomic composition describes the content of the five atoms (carbon, nitrogen, hydrogen, sulfur, and oxygen) constituting amino acids which are the components of a particular protein. The atomic composition of a protein can be used to calculate the precise molecular weight of the entire protein.

The amino acid count is the sum of all amino acids in a molecule.

The amino acid frequency is the number of each of the 20 amino acids as a proportion of the total number of amino acids.

Feature Selection procedure

It is likely that only a few of the parameters are important, so it is necessary to be able to eliminate parameters which have no influence on thermo- or halostability. Feature selection is used to distinguish informative attributes from less informative ones. In order to investigate features that affect the thermostability and alkaline activity of an enzyme, Clementine SPSS v.11.0 (http://www-01.ibm.com/software/analytics/spss/products/modeler/) was used.

Feature selection algorithms are typically divided into three types: Feature Screening, Ranking and Subset Selection. Feature Screening removes unimportant, problematic predictors and cases. Feature Ranking sorts the remaining predictors and assigns ranks based on importance. The measure used to rank importance depends on whether the predictors and the targets are all categorical, numeric ranges, or a mix of range and categorical. Subset Selection identifies the important subsets in the features.

Utilized Algorithm

The Feature Selection algorithm can be used to identify the fields that are most important for a given analysis. Since all predictors and targets have continuous P values based on the F statistic used, the idea is to perform a one-way ANOVA F test for each continuous predictor.

F-test

A hypothesis test examines the ratio of two variances to determine their equality. Typically one-tailed

F-tests, refer to the F-distribution. An F-test evaluates whether the observed statistic exceeds a critical value from a distribution. If the observed F-statistic exceeds the critical value, the null hypothesis is rejected:

$$F = \frac{S_1^2}{S_2^2}$$

S_1^2 = the variance of the first sample
S_2^2 = the variance of the second sample
Let's consider the prediction of a continuous outcome *y*. The Pearson correlation coefficient is defined as:

$$R(i) = \frac{\mathrm{cov}(Xi, Y)}{\sqrt{\mathrm{var}(Xi)\,\mathrm{var}(Y)}} \quad (1)$$

the estimate of $R(i)$ is given by:

$$R(i) = \frac{\sum_{k=1}^{m}(x_{k,i} - \bar{x})\,(y_k - \bar{y})}{\sqrt{\sum_{k-1}^{m}(x_{k,i} - \bar{x}_i)^2 \sum_{k=1}^{m}(y_k - \bar{y})^2}} \quad (2)$$

where the bar notation stands for an average over the index *k*. Although $R(i)$ is derived from $R(i)$, it may be used without assuming that the input values are realizations of a random variable. In a linear regression, the coefficient of determination, which is the square of $R(i)$, represents the fraction of the total variance around the mean value *y* that is explained by the linear relation between *x* and *y*. Therefore, using $R(i)$ as a variable ranking criterion enforces a ranking according to the goodness of linear fit of the individual variables. Correlation criteria such as $R(i)$ can only detect linear dependencies between the variable and the target. A simple way of lifting this restriction is to make a non-linear fit of the target with single variables and rank according to the goodness of fit.

Modeling thermostability or alkaline activity based on the structural amino acid attributes by multiple linear regression

Understanding the relationship between an enzyme's important features and its thermostability or alkaline activity is essential. To achieve this, linear correlation coefficients (r) between each of the 74 features and proteins were calculated using SPSS 14.0 software.

All 81 proteins were divided into two groups: thermostable and non-thermostable. The thermostable group included enzymes which can be active in temperatures higher than 80°C. The non-thermostable group contained enzymes which are only active in temperatures lower than 80°C. A *t*-test was used to compare the means of these two groups. The *t*-test compared the actual difference between two means in relation to the variation in the data. Two data sets were classified (<80°C and >80°C), and each one was characterized by its mean. A *t*-test was used to determine whether each feature in the two data sets was significant.

Taking into account the industrial demand for enzymes with high levels of halo- and thermostability, this study may provide the required knowledge for the simultaneous engineering of proteins in hot and alkaline conditions. To examine the effectiveness of predictors in practical applications, the following three cross-validation methods are often used: independent dataset test, subsampling or *k*-fold crossover test, and jackknife test. In this study, we used 2-fold cross-validation. Cross validation is a commonly used method for evaluating classifier methods, as sets of proteins used for training and testing are mutually exclusive. To this end, we first divided the records (xylanase accession numbers) into two classes: xylanases active in temperatures than 50°C and xylanases active in temperatures lower than 50°C. The records corresponding to activities equal to 50°C were omitted. The same process was performed for pH, and pH = 7 was removed from the cross validation. Then, selected records were randomly divided into 10 (nearly equal in number) sets. Nine parts were used for training the prediction algorithm (regression model) and the 10^{th} for testing. Then, in the next run, another part set was used as a testing set and the other 9 parts as training sets. The process was repeated twice and the accuracy levels for true, false and total accuracy were calculated. The final accuracy reported was the average of the accuracy levels of two tests (runs).

Increasing the accuracy of regression models

To increase the accuracy of predictive models, the following procedures and conditions were examined: 1) removing and adding features (predictors) to the model for the purpose of identifying a useful subset of the features using standard stepwise (adds and removes terms), forward selection (adds terms), and backward elimination (removes terms, and 2) standardization or non-standardization of features (continuous predictors). A stepwise procedure starts with an empty model and then adds or removes a feature for each step. The alpha coefficient, to enter or remove, was set to 0.15 in a step-

wise procedure. A forward selection procedure starts with an empty model and adds the most significant feature for each step. The alpha coefficient to enter was set to 0.25 in the forward selection. A backward elimination procedure starts with all potential terms in the model and removes the least significant term for each step. The alpha coefficient to remove was set to 0.1 in the backward elimination. All procedures (stepwise/forward/backward) were tested under two conditions of standardization/non-standardization of the features (continuous predictors). Standardizing the continuous predictors can improve the interpretation of the model. The continuous attributes were standardized by subtracting the mean and dividing by the standard deviation. Subtracting the mean helps to reduce multicollinearity, which improves the precision of the coefficient estimates. In particular, this method is helpful when the model contains highly correlated predictors, higher-order terms, and interaction terms. Dividing the values of the attributes by the standard deviation allows the comparison of the size of the coefficients in a comparable scale. This approach helps to determine which predictors have a larger effect through controlling for differences in scale.

In all of the above examined algorithms, R-square, the coefficient of determination, was determined by a comparison of the regression sum of squares with the total sum of squares. R-square determines the total variation that is explained by the regression model. The larger this value the better the relationship formula, explaining thermostability/pH as a function of structural amino acid attributes. The confidence level for all intervals was adjusted to 95% in all cases. Adjusted sums of squares (type III) were used as this does not depend on the order of the factors that are entered into the model. The analysis was performed using Minitab 17 (http://www.minitab.com/en-us/products/minitab/).

Results

The studied xylanases were active between 20 and 102°C and the median temperature for activation was 65°C. These proteins were stable at a pH ranging between 2 to 11. The median pH was 6. Using the sequences obtained from the Swiss-Prot and TrEMBL databases and publications, 74 characteristic features were generated. These features relate to the primary structure of the proteins.

Feature selection

An ANOVA F-test was used for carrying out feature selection on the datasets. There were 71 sequence data sets for temperature and 85 for pH. In the present study, we concentrated on a computational analysis of extracted features. According to the performed feature selection, features were classified as important, unimportant or marginal. After investigation, some attributes were identified as important features. These features play a critical role in the thermostability and alkalophilicity of an enzyme. The frequency and count of some amino acids are important features for the activation of the enzyme at high pHs (see supplementary Table 1 for details). The same procedure also indicated that the frequency and count of some amino acids and atoms play a role in the thermostability of this enzyme (see supplementary Table 2 for details).

Modeling

Linear correlation coefficient (r) was calculated to estimate the relationship between important features, pH and optimum temperature data.

The analysis of thermostability data of xylanases showed a positive correlation between thermostability and the following protein features: length, weight, non-reduced Cys absorption coefficient, reduced Cys absorption coefficient, hydrophobic residue count, negatively charged residues, positively charged residues, Glu, Phe, Ile, Lys Pro, Arg, Val, Trp. A negative correlation was found between temperature and the oxygen and Cys counts (see supplementary Table 3 for details).

The optimum pH values indicate that the His and Arg counts, frequency for Asp, Glu, His, hydrogen, negatively charged residues, positively charged residues, Lys, Leu, Arg are positively correlated with optimum pH. In addition, hydrophilic residue, oxygen, Ser, Thr, and Tyr revealed a negative correlation (see supplementary Table 4 for details).

T-tests confirmed our previous results and showed a significant difference between the means of the discovered features in the two thermostability data sets (< 80°C and > 80°C) (see supplementary Table 5 for details).

A multiple linear regression was used for derivation of optimum temperature and optimum pH as a dependent variable and for other features as an independent variable. The final model from the stepwise analysis con-

Table 1. Multiple linear regression for modeling optimum xylanase temperature and pH (as dependent variables) based on structural amino acid features

Dependent variable	B	Features (Independent)	Beta	Std error	Sig.
Optimum temperature	53.648	Reduced cysteine extinction coefficient	0.260	0.057	0.000
		Count of sulfur	−1.121	0.314	0.000
		Count of Glutamic acid	−0.763	0.249	0.001
		Count of Arginine	0.637	0.240	0.003
		R-square = 37.2%			
Optimum pH	6.411	Frequency of Serine	−0.473	4.449	−5.053
		Isoelectric point	0.208	0.092	2.223
		R-square = 29.2%			

tains only four features for optimum temperature and two features for optimum pH (Table 1). The prediction accuracy levels for models were evaluated by 10-fold cross validation. An accuracy of 92% for optimum temperature and an accuracy of 86% for optimum pH were achieved.

As presented in Table 2, application of forward selection and backward elimination increased the accuracy of temperature prediction to 57.76% and 85.12%, respectively. Also, forward selection and backward elimination noticeably increased the accuracy of the pH regression model to 41.37% and 7784% (Table 3). Feature standardization provided the possibility of a more accurate comparison of structural feature importance. Q0H3C1 accession was a false positive protein which had an optimum temperature of 20°C, but the model predicted the temperature of 66.55°C. On the other hand, P23557 accession had an optimum temperature of 75°C, but the model fitted the temperature of 44.37°C. In the case of optimum pH, P96174 had the optimum basic pH of 9 but the model fitted the acidic pH of 5.82. Conversely, the P48793 protein had the pH of 5 but the value of 6.52 was predicted by the model.

Laboratory errors in measuring the temperature and pH might be one of the reasons for the above mentioned false-negative and false-positive cases. It is also probable that these proteins follow other mechanisms in thermostability and alkalophilicity which have not been captured by the developed models in this study.

Discussion

Bioinformatics can save months of work in the lab at the cost of a few hours of work with a computer. Bioinformatic studies help researchers in protein engineering. Feature selection is one of the most important and frequently used techniques in the data pre-processing used in data mining (Liu et al., 2010).

Features which affect the optimum pH for xylanase activities

The present data suggest that there are 18 important xylanase (protein) features which have an influence on optimum pH for enzymatic activity.

Frequency of occurrence of hydrophilic residues, such as Serine, Threonine, Tyrosine and Glutamine

Each amino acid has been classified within a specific group based on its biophysical and biochemical properties. Consequently, when a special group is defined as an important feature, the members of that group are also defined as important features and vice versa. For instance, it can be seen that the frequency of appearance of hydrophilic residues is an important feature (value = 1). Hence, the frequency of the members of this group, such as, Ser (value = 1), Thr (value = 0.996), Tyr (value = 0.993) and Gln (value = 0.972), are also considered important features. It is important to note that pH was negatively correlated with these features. This can be explained by the hydroxyl group in their structures, which enables them to be stable in low pH conditions in which hydrogen ion concentration is high. More importantly, the results revealed that the negative correlation was statistically significant ($P < 0.05$). The frequency of positively charged residues such as, Arg, and Lys (value = 1) has an effective role on pH and the activity of the enzyme. It seems that their positive charge, which is retai-

Table 2. Optimization of regression model for xylanase thermostability (as dependent variables) based on structural amino acid attributes

Stepwise selection			Stepwise selection with standardization		
Term	Coefficient	*P*-value	Term	Coefficient	*P*-value
Constant	51.61	0.000	Constant	62.24	0.000
Non-reduced cysteines extinction	0.2512	0.002	Non-reduced cysteines extinction	14.04	0.002
Count of sulfur	−1.127	0.001	Count of sulfur	−9.12	0.001
Count of sulfur	−0.351	0.121	Count of sulfur	−8.98	0.121
Count of Pro	0.719	0.023	Count of Pro	10.28	0.023
Count of Gln	−0.68	0.026	Count of Gln	−8.5	0.026
Count of Arg	0.514	0.044	Count of Arg	5.97	0.044
R-square = 41.23%			R-square = 41.23%		
Regression equation: optimum temperature = 51.61; +0.2512 – non-reduced cysteines extinction; −1.127 – count of sulfur; −0.351 – count of Asp; +0.719 – count of Pro; −0.680 – count of Gln; +0.514 – count of Arg					
Forward selection			Forward selection with standardization		
Constant	143.6	0.000	Constant	62.24	0.000
Non-reduced cysteines extinction	0.183	0.075	Non-reduced cysteines extinction	10.21	0.075
Count of sulfur	−0.931	0.034	Count of sulfur	−7.53	0.034
Count of hydrophilic of residues	0.417	0.003	Count of hydrophilic of residues	39.1	0.003
Frequency of other residues	−178.1	0.022	Frequency of other residues	−11.77	0.022
Count of other charged residues	0.0207	0.153	Count of other charged residues	4.4	0.153
Count of Ala	−0.47	0.007	Count of Ala	−12.54	0.007
Count of Glu	−0.079	0.759	Count of Glu	−1.73	0.759
Count of Gln	−1.239	0.003	Count of Gln	−15.48	0.003
Count of Arg	0.917	0.008	Count of Arg	10.65	0.008
Count of Ser	−1.592	0.000	Count of Ser	−35.77	0.000
Count of Val	0.492	0.128	Count of Val	11.92	0.128
Frequency of Gly	−171	0.095	Frequency of Gly	−6.09	0.095
Frequency of Met	−493	0.083	Frequency of Met	−4.01	0.083
Frequency of Asn	−181	0.147	Frequency of Asn	−3.72	0.147
Frequency of Tyr	−327	0.058	Frequency of Tyr	−6.17	0.058
R-square = 57.76%			R-square = 57.76%		
Regression equation: optimum temperature = 143.6; +0.183 – non-reduced cysteines extinction; −0.931 – count of sulfur; +0.417 – count of hydrophilic residues; −178.1 – frequency of other residues; +0.0207 – count of other charged residues; −0.470 – count of Ala; −0.079 – count of Glu; −1.239 – count of Gln; +0.917 – count of Arg; −1.592 – count of Ser; +0.492 – count of Val; −171 – frequency of Gly; −493 – frequency of Met; −181 – frequency of Asn; −327 – frequency of Tyr					
Backward elimination			Backward elimination with standardization		
Term	Coefficient	*P*-value	Term	Coefficient	*P*-value
Constant	4379	0.000	Constant	62.24	0.000
Length	75.2	0.000	Length	23490	0.000
Weight (Kda)	−494	0.000	Weight (Kda)	−17214	0.000
Half-life in mammals (hours)	2.521	0.003	Half-life in mammals (hours)	17.12	0.003
Half-life in yeast (> hours)	−4.78	0.001	Half-life in yeast (> hours)	−15.95	0.001
Half-life in *E. coli* (> hours)	−31.9	0.018	Half-life in *E. coli* (> hours)	-38	0.018

Non-reduced cysteines extinction	−516.3	0.000	Non-reduced cysteines extinction	−28848	0.000
Non-reduced cysteines absorption	−9248	0.060	Non-reduced cysteines absorption	−5263	0.060
Reduced cysteines extinction coefficient	523.5	0.000	Reduced cysteines extinction coefficient	29233	0.000
Reduced cysteines absorption	9322	0.058	Reduced cysteines absorption	5302	0.058
Count of carbon	0.0407	0.030	Count of carbon	4.77	0.030
Count of hydrogen	−0.785	0.004	Count of hydrogen	−47.9	0.004
Frequency of sulfur	36681	0.000	Frequency of sulfur	29.52	0.000
Frequency of carbon	−11784	0.000	Frequency of carbon	−52.3	0.000
Count of hydrophobic residues	−23.85	0.000	Count of hydrophobic residues	−3499	0.000
Count of negatively other charged res	−11.28	0.000	Count of negatively other charged	−506	0.000
Frequency of other charged res	−971	0.006	Frequency of other charged res	−58.2	0.006
Count of Ala	−17.15	0.000	Count of Ala	−457.4	0.000
Count of Asp	−9.29	0.000	Count of Asp	−237.8	0.000
Count of Phe	14.37	0.001	Count of Phe	197.5	0.001
Count of Gly	−21.67	0.000	Count of Gly	−534	0.000
Count of Iie	4.65	0.004	Count of Iie	95.1	0.004
Count of Lys	−10.38	0.000	Count of Lys	−247.9	0.000
Count of Leu	3.97	0.021	Count of Leu	91.9	0.021
Count of Met	17.84	0.000	Count of Met	115.7	0.000
Count of Asn	−18.27	0.000	Count of Asn	−409.9	0.000
Count of Gln	−14.88	0.000	Count of Gln	−186	0.000
Count of Ser	−33.74	0.000	Count of Ser	−758	0.000
Count of Thr	−26.13	0.000	Count of Thr	−660	0.000
Frequency of Ala	736	0.017	Frequency of Ala	40.2	0.017
Frequency of Cys	6002	0.023	Frequency of Cys	48	0.023
Frequency of Glu	−747	0.097	Frequency of Glu	−17.7	0.097
Frequency of Phe	6116	0.000	Frequency of Phe	66.8	0.000
Frequency of His	−1824	0.001	Frequency of His	−19.92	0.001
Frequency of Iie	805	0.061	Frequency of Iie	13.44	0.061
Frequency of Lys	−1134	0.008	Frequency of Lys	−28.3	0.008
Frequency of Leu	1016	0.038	Frequency of Leu	21.23	0.038
Frequency of Met	−4908	0.000	Frequency of Met	−39.98	0.000
Frequency of Gln	1839	0.004	Frequency of Gln	22	0.004
Frequency of Ser	620	0.064	Frequency of Ser	18.25	0.064
Frequency of Thr	1044	0.002	Frequency of Thr	32.21	0.002
R-square = 85.12%			R-square = 85.12%		

Regression equation: optimum temperature = 4379; +75.2 – length; −494 – weight *t* (Kda); +2.521 – half-life in mammals (hours); −4.78 – half-life in yeast (> hours); −31.9 – half-life in *E. coli* (> hours); −516.3 – non-reduced cysteines extinction; −9248 – non-reduced cysteines absorption; +523.5 – reduced cysteines extinction coe; +9322 – reduced cysteines absorption; +0.0407 – count of carbon; −0.785 – count of hydrogen; +6681 – frequency of sulfur; −11784 – frequency of carbon; −23.85 – count of hydrophobic residues; −11.28 – count of negatively charged res; −971 – frequency of other charged res; −17.15 – count of Ala; −9.29 – count of Asp; +14.37 – count of Phe; −21.67 – count of Gly; +4.65 – count of Iie; −10.38 – count of Lys; +3.97 – count of Leu; +17.84 – count of Met; −18.27 – count of Asn; −14.88 – count of Gln; −33.74 – count of Ser; −26.13 – count of Thr; +736 – frequency of Ala; +6002 – frequency of Cys; −747 – frequency of Glu; +6116 – frequency of Phe; −1824 – frequency of His; +805 – frequency of Iie; −1134 – frequency of Lys; +1016 – frequency of Leu; −4908 – frequency of Met; +1839 – frequency of Gln; +620 – frequency of Ser; +1044 – frequency of Thr

Table 3. Optimization of regression model for optimum pH (as dependent variables) based on structural amino acid attributes

Stepwise selection			Stepwise selection with standardization		
Term	Coefficient	*P*-value	Term	Coefficient	*P*-value
Constant	0.710	0.745	Constant	5.828	0.000
Isoelectric point	0.599	0.000	Isoelectric point	0.999	0.000
Count of carbon	0.002	0.086	Count of carbon	0.263	0.086
Frequency of negatively charged	30.900	0.003	Frequency of negatively charged	1.122	0.003
Frequency of Lys	−21.200	0.049	Frequency of Lys	−0.537	0.049
Frequency of Ser	−21.910	0.007	Frequency of Ser	−0.759	0.007
Frequency of Thr	12.600	0.112	Frequency of Thr	0.400	0.112
R-square = 39.9%			R-square = 39.1%		
Regression equation: optimum PH = 0.71; +0.599 – isoelectric point; +0.001500 – count of carbon; +30.9 – frequency of negatively charged; −21.2 – frequency of Lys; −21.91 – frequency of Ser; +12.60 – frequency of Thr					
Forward selection			Forward selection with standardization		
Term	Coefficient	*P*-value	Term	Coefficient	*P*-value
Constant	−2.48	0.397	Constant	5.828	0.000
Isoelectric point	0.659	0.000	Isoelectric point	1.1	0.000
Count of carbon	0.001	0.183	Count of carbon	0.214	0.183
Frequency of negatively charged	39.2	0.001	Frequency of negatively charged	1.425	0.001
Frequency of Gly	8.59	0.231	Frequency of Gly	0.317	0.231
Frequency of Lys	−16.30	0.142	Frequency of Lys	−0.414	0.142
Frequency of Gln	15.6	0.124	Frequency of Gln	0.251	0.124
Frequency of Ser	−17.80	0.035	Frequency of Ser	−0.617	0.035
Frequency of Thr	12.97	0.110	Frequency of Thr	0.411	0.110
R-square = 41.37%			R-square = 41.37%		
Regression equation: optimum PH = −2.48; +0.659 – isoelectric point; +0.001222 – count of carbon; +39.2 – frequency of negatively charged; +8.59 – frequency of Gly; −16.3 – frequency of Lys; +15.6 – frequency of Gln; −17.80 – frequency of Ser; +12.97 – frequency of Thr					
Backward elimination			Backward elimination with standardization		
Constant	−402.6	0.000	Constant	5.828	0.000
Length	−212.5	0.010	Length	−68831	0.010
Weight (Kda)	1554	0.010	Weight (Kda)	56203	0.010
Isoelectric point	0.78	0.004	Isoelectric point	1.301	0.004
Aliphatic index	0.3622	0.001	Aliphatic index	4.98	0.001
Half-life in mammals (hours)	−0.1165	0.022	Half-life in mammals (hours)	−0.893	0.022
Half-life in yeast (> hours)	0.1425	0.064	Half-life in yeast (> hours)	0.6	0.064
Non-reduced cysteines extinction	1045	0.010	Non-reduced cysteines extinction	70318	0.010
Non-reduced cysteines absorption	−849	0.047	Non-reduced cysteines absorption	−580	0.047
Reduced cysteines extinction coefficient	−1069	0.010	Reduced cysteines extinction coefficient	−71876	0.010
Reduced cysteines absorption	848	0.047	Reduced cysteines absorption	580	0.047
Count of oxygen	0.00194	0.100	Count of oxygen	0.514	0.100
Frequency of carbon	863	0.002	Frequency of carbon	3.94	0.002
Count of hydrophobic residues	58.6	0.010	Count of hydrophobic residues	8876	0.010
Count of hydrophilic residues	−10.41	0.007	Count of hydrophilic residues	−1042	0.007

Frequency of hydrophobic residues	316.4	0.000	Frequency of hydrophobic residues	14.05	0.000
Frequency of hydrophilic residues	320	0.000	Frequency of hydrophilic residues	27.58	0.000
Count of negatively charged residues	11.9	0.014	Count of negatively charged residues	546	0.014
Count of positively charged residues	−30.1	0.009	Count of positively charged residues	−1009	0.009
Frequency of other charged residues	−212.9	0.002	Frequency of other charged residues	−13.05	0.002
Count of Ala	43.5	0.010	Count of Ala	1216	0.010
Count of Asp	21.78	0.010	Count of Asp	581	0.010
Count of Phe	−75	0.010	Count of Phe	−994	0.010
Count of Gly	65.3	0.010	Count of Gly	1570	0.010
Count of Iie	−21.92	0.010	Count of Iie	−464	0.010
Count of Lys	43.7	0.010	Count of Lys	1046	0.010
Count of Leu	−22.13	0.009	Count of Leu	−531	0.009
Count of Met	−50.2	0.009	Count of Met	−334	0.009
Count of Asn	45.7	0.010	Count of Asn	1404	0.010
Count of Pro	3.06	0.010	Count of Pro	43.9	0.010
Count of Gln	23.9	0.010	Count of Gln	419	0.010
Count of Ser	87.6	0.010	Count of Ser	1990	0.010
Count of Thr	65.8	0.010	Count of Thr	1610	0.010
Frequency of Ala	−45.3	0.027	Frequency of Ala	−2.32	0.027
Frequency of Cys	522	0.036	Frequency of Cys	2.75	0.036
Frequency of Asp	139.6	0.009	Frequency of Asp	2.74	0.009
Frequency of Glu	94.6	0.072	Frequency of Glu	2.1	0.072
Frequency of Iie	−108.9	0.024	Frequency of Iie	−1.884	0.024
Frequency of Arg	153.5	0.001	Frequency of Arg	2.274	0.001
Frequency of Val	−143.7	0.000	Frequency of Val	−2.511	0.000
Frequency of Tyr	−151.9	0.037	Frequency of Tyr	−2.95	0.037
R-square = 77.84%			R-square = 77.84%		

Regression equation: optimum PH = −402.6; −212.5 – length; +1554 – weight (Kda); +0.780 – isoelectric point; +0.3622 – aliphatic index; −0.1165 – half-life in mammals (hours); +0.1425 – half-life in yeast (> hours); +1045 – non-reduced cysteines extinction; −849 – non-reduced cysteines absorption; −1069 – reduced cysteines extinction coe; +848 – reduced cysteines absorption; +0.00194 – count of oxygen; +863 – frequency of carbon; +58.6 – count of hydrophobic residues; −10.41 – count of hydrophilic residues; +316.4 – frequency of hydrophobic residue; +320.0 – frequency of hydrophilic residue; +11.90 – count of negatively charged res; −30.1 – count of positively charged res; −212.9 – frequency of other charged res; +43.5 – count of Ala; +21.78 – count of Asp; −75.0 – count of Phe; +65.3 – count of Gly; −21.92 – count of Iie; +43.7 – count of Lys; −22.13 – count of Leu; −50.2 – count of Met; +45.7 – count of Asn; +3.06 – count of Pro; +23.90 – count of Gln; +87.6 – count of Ser; +65.8 – count of Thr; −45.3 – frequency of Ala; +522 – frequency of Cys; +139.6 – frequency of Asp; +94.6 – frequency of Glu; −108.9 – frequency of Iie; +153.5 – frequency of Arg; −143.7 – frequency of Val; −151.9 – frequency of Tyr

ned in alkaline conditions, helps the molecules become more stable at high pHs. As a result, it is shown that the frequency of Arg (value = 1) and Lys (value = 0.997) and the Arg count (value = 0.982) are recognized as important features. Statistical analysis indicated a positive correlation between pH and these features. Therefore, an increase in these features will increase the stability of xylanase in alkaline conditions. This was also considered to have a stabilizing effect on the enzyme. Arg was shown to retain a positive charge under high alkaline conditions and form ion pairs with neighboring residues (Shirai et al., 1997).

Frequency of negatively charged residues (Asp and Glu)

Our results revealed that the frequency of Asp and Glu are important features regarding enzyme pH stability. Previous studies have shown that, taken together, protein surfaces rich in acidic residues may help the pro-

tein carry out its enzymatic activity in alkaline pH environments and protect the protein core from an OH attack (Manikandan et al., 2009). In our study, feature selection along with a positive correlation confirmed this. Russell and Fersht (1987) described how an increase in the negative charge on the surface of the model enzyme subtilisin raised the pKa values of acidic groups and how, with an increased pKa in the catalytic acidic amino acid, the enzyme can shift its optimum pH to alkaline values (Russell and Fersht, 1987).

Aliphatic Index, frequency of Leu and Val

The aliphatic index of a protein is a measure of the relative volume occupied by the aliphatic side chain of the following amino acids: Ala, Val, Leu and Ile. As previously mentioned, the output shows them as important features with the value = 0.958, 1 and 0.975, respectively. The aliphatic index and frequency of Leu were positively correlated. However, the frequency of Val was negatively correlated with the aliphatic index. It is clear that further research is needed to obtain better definitions for the relationships between these characteristics, not only for xylanases, but also for other enzymes.

Effective features of the thermostability of xylanases

The feature selection performed on the collected data marked 21 features as important attributes contributing to the thermostable conformation of the protein.

Hydrophobic residue count

The hydrophobic contact between the monomers of the enzyme may affect its thermostability. The results of our study indicate that the counts of: Ile (value = 0.993), Trp (value = 0.992), Pro (value = 0.992), Phe (value = 0.989), Val (value = 0.988) as well as the hydrophobic residue counts (value = 0.966) are important features. Moreover, they have a positive correlation with the thermostability of the enzyme. These features cause an increase in the hydrophobic contacts between monomers and consequently the enzyme can become stable at high temperatures. In previous studies, it was noted that the major driving force behind thermostability is the increased hydrophobicity between monomers (Miyazaki et al., 2006).

Positively (Arg and Lys) and negatively charged residue counts (Asp and Glu)

Research has been performed with results indicating that salt bridges between oppositely charged groups are an important aspect in protein structure and stability (Anderson et al., 1990). Besides, based on Elcock's experiment in 1998, this type of ion pair interaction leads to stability at high temperatures. Salt bridges can be expected to have thermal stability and maintain their activity at elevated temperatures (Natesh et al., 1999). Our study also confirms these outputs. As indicated by our results, counts of: Glu (value = 0.995), positively charged residues (value = 0.976), Lys (value = 0.957) and Arg (value = 0.978) have been identified as important features. In accordance with the biochemical properties of these amino acids, Glu is negatively charged because of its possession of two extra carboxyl groups. On the other hand, Lys and Arg are positively charged because of an extra amino group – NH_2 in their structures. Accordingly, two charged groups produce salt bridges (charge-assisted hydrogen bonding) when bound together. This kind of ionic-bond has a key role in the stability of the enzyme at high temperatures.

Retained features such as weight and length were also recognized as important features. Indeed, a statistical analysis showed a positive correlation between those features and thermostability. While the length and the weight of the protein were defined as important features, they are not suitable characteristics in industry.

Among all the important features which were determined for the thermostability and alkalophlicity of xylanase, the Arg count was responsible for both characteristics. This reveals a critical feature which plays a key role in molding the activities of thermotolerant and alkalophilic enzymes. As a result, if the number of Args increases, a thermo-alkalophilic enzyme can be produced.

We suggest that in future studies nonlinear models as well as novel machine learning models such as decision trees, bayesian models, and neural networks be employed for fitting the obtained data. It is especially important for the description of pH dependency, as the presented 2-variable linear model is definitely insufficient. As the high performance of the feature selection algorithm in the coordinated analysis of thermostability and alkalophilicity was proved in this study, novel data mining models have the potential to open a new way forward in this field.

Conclusions

Use of thermophilic and alkalophilic xylanases is desirable and appealing in industrial applications. It is

worth mentioning that there are only a few microorganisms that are able to secrete a thermophilic xylanase. Efforts to obtain/acquire a thermophilic and alkalophilic enzyme are still being continued. Scientists are trying to achieve this goal by using PCR prone or directed mutations to produce amino acid substitutions. However, to date there is no evidence about the exact substitutions of amino acids which would result in higher xylanase stability at both high temperatures and pH levels. Thus, we should enhance our knowledge about the effective attributes which are important for obtaining the desired enzyme. Bioinformatic methods are prone to guide the research in a specified direction. In this article, we first investigated the attributes of 101 data sets by extracting features from the proteins and performing a feature selection. Feature selection allowed the variable set to be reduced in size and created a more manageable set of attributes. It enabled us to center our attention on the informative features that play a key role on the stability and activity of the enzyme at elevated temperatures and pH. The results of the statistical analysis showed a relationship between these important features and stability.

References

Anderson D.E., Becktel W.J., Dahlquist F.W. (1990) *pH-induced denaturation of proteins: a single salt bridge contributes 3-5 kcal/mol to the free energy of folding of T4 lysozyme*. Biochemistry 29: 2403-2408.

Arase A., Yomo T., Urabe I., Hata Y., Katsube Y., Okada H. (1993) *Stabilization of xylanase by random mutagenesis*. FEBS Lett. 316: 123-127.

Ashrafi E., Alemzadeh A., Ebrahimi M., Ebrahimie E., Dadkhodaei N. (2011) *Amino acid features of P1B-ATPase heavy metal transporters enabling small numbers of organisms to cope with heavy metal pollution*. Bioinformatics Biol. Ins. 5: 5-9.

Bakhtiarizadeh M.R., Moradi-Shahrbabak M., Ebrahimi M., Ebrahimie E. (2014) *Neural network and SVM classifiers accurately predict lipid binding proteins, irrespective of sequence homology*. J. Theor Biol. 356: 213-222.

Bijanzadeh E., Emam Y., Ebrahimie E. (2010) *Determining the most important features contributing to wheat grain yield using supervised feature selection model.* Austral. J. Crop Sci. 4: 402-407.

Chen Y.-L., Tang T.-Y., Cheng K.-J. (2001) *Directed evolution to produce an alkalophilic variant from a Neocallimastix patriciarum xylanase*. Can. J. Microbiol. 47: 1088-1094.

Cheng H.-L., Wang P.-M., Chen Y.-C., Yang S.-S., Chen Y.-C. (2008) *Cloning, characterization and phylogenetic relationships of a endoxylanase-encoding gene from Streptomyces thermonitrificans NTU-88*. Bioresour. Technol. 99: 227-231.

Collins T., Gerday C., Feller G. (2006) X*ylanases, xylanase families and extremophilic xylanases*. Fems Microbiol. Rev. 29: 3-23.

Decelle B., Tsang A., Storms R.K. (2004) *Cloning, functional expression and characterization of three Phanerochaete chrysosporium endo-1, 4ß-xylanases*. Curr. Genet. 46: 166-175.

Ebrahimi M., Aghagolzadeh P., Shamabadi N., Tahmasebi A., Alsharifi M., Adelson D.L., Hemmatzadeh F., Ebrahimie E. (2014) *Understanding the undelaying mechanism of HA-subtyping in the level of physic-chemical characteristics of protein.* PloS one 9: e96984.

Ebrahimi M., Ebrahimie E. (2010) *Sequence-based prediction of enzyme thermostability through bioinformatics algorithms*. Curr. Bioinform. 5: 195-203.

Ebrahimi M., Ebrahimie E., Shamabadi N., Ebrahimi M. (2010) *Are there any differences between features of proteins expressed in malignant and benign breast cancers?* J. Res. Med. Sci. 15: 299-309.

Ebrahimi E., Ebrahimi M., Sarvestani N.R., Ebrahimi M. (2011) *Protein attributes contribute to halo-stability, bioinformatics approach.* Saline Syst. 7: 1-14.

Krämer N., Schäfer J., Boulesteix A.-L. (2009) *Regularized estimation of large-scale gene association networks using graphical Gaussian models*. BMC Bioinform. 10: 384.

Kulkarni N., Shendye A., Rao M. (2006) *Molecular and biotechnological aspects of xylanases*. FEMS Microbiol. Rev. 23: 411-456.

Lakizadeh A., Agha-Golzadeh P., Ebrahimi M., Ebrahimie E., Ebrahimi M. (2011) *Engineering thermostable enzymes*. Adv. Stud. Biol. 3: 63-78.

Liu H., Motoda H., Setiono R., Zhao Z. (2010) *Feature selection: An ever evolving frontier in data mining*. Proc. The Fourth Workshop on Feature Selection in Data Mining, pp. 4-13.

Manikandan K., Bhardwaj A., Gupta N., Lokanath N.K, Ghosh A., Reddy V.S., Ramakumar S. (2009) *Crystal structures of native and xylosaccharide bound alkali thermostable xylanase from an alkalophilic Bacillus sp. NG 27: Structural insights into alkalophilicity and implications for adaptation to polyextreme conditions*. Protein Sci. 15: 1951-1960.

Miyazaki K., Takenouchi M., Kondo H., Noro N., Suzuki M., Tsuda S. (2006) *Thermal stabilization of Bacillus subtilis family-11 xylanase by directed evolution*. J. Biol. Chem. 281: 10236-10242.

Natesh R., Bhanumoorthy P., Vithayathil P., Sekar K., Ramakumar S., Viswamitra M. (1999) *Crystal structure at 1.8 Å resolution and proposed amino acid sequence of a thermostable xylanase from Thermoascus aurantiacus*. J. Mol. Biol. 288: 999-1012.

Russell A.J., Fersht A.R. (1987) *Rational modification of enzyme catalysis by engineering surface charge*. Nature 328: 496.

Shekoofa A., Emam Y., Ebrahimi M., Ebrahimie E. (2011) *Application of supervised feature selection methods to de-*

fine the most important traits affecting maximum kernel water content in maize. Austral. J. Crop Sci. 5: 162-168.

Shirai T., Suzuki A., Yamane T., Ashida T., Kobayashi T., Hitomi J., Ito S. (1997) *High-resolution crystal structure of M-protease: phylogeny aided analysis of the high-alkaline adaptation mechanism*. Protein Eng. 10: 627-634.

Shoham Y., Schwartz Z., Khasin A., Gat O., Zosim Z., Rosenberg E. (1992) *Delignification of wood pulp by a thermostable xylanase from Bacillus stearothermophilus strain T-6*. Biodegradation 3: 207-218.

Xie H., Flint J., Vardakou M., Lakey J.H., Lewis R.J., Gilbert H.J., Dumon C. (2006) *Probing the structural basis for the difference in thermostability displayed by family 10 xylanases*. J. Mol. Biol. 360: 157-167.

Zinati Z., Zamansani F., Kayvanjoo A.H., Ebrahimi M., Ebrahimi M., Ebrahimie E., Mohammadi Dehcheshmeh M. (2014) *New layers in understanding and predicting α-linolenic acid content in plants using amino acid characteristics of omega-3 fatty acid desaturase*. Comput. Biol. Med. 54: 14-23.