2009 International Conference on Machine Learning and Computing
IPCSIT vol.3 (2011) © (2011) IACSIT Press, Singapore

Using neural networks expert system to predict protein thermostability

Esmaeil Ebrahimie [1*], Mansour Ebrahimi [2], Tahereh Deihimi [3], Mahdi Ebrahimi [4]

[1*] Department of Crop Production & Plant Breeding, College of Agriculture, Shiraz University, Shiraz, Iran,
[2] Bioinformatics Research Group, Green Research Center, Qom University, Qom, Iran
[3] Plant Biotechnology Center, College of Agriculture, Shiraz University, Shiraz, Iran
[4] Department of Information Technology, International University in Germany, Bruchsal, Germany

Abstract. Some biological or chemical reactions need to be performed at high temperatures to decrease reaction time. However, many proteins are not very stable when heated. Research is needed that helps proteins to remain active and stable at high temperatures overcoming many limitations to their industrial applications. Recently we have shown that some structural features of proteins are related to the thermostability. However, a general model of how these features are related to thermostability remained a scientific challenge. Our goal was to develop a predictive model for thermostability based on sequence and structural features by using neural networks modeling. Amino acid sequences of 4636 proteins, including 3069 mesophilic and 1294 thermophilic, were extracted from databases, 74 protein features calculated and feature selection algorithm applied to determine the most important features contributing to thermal stability. Six different neural network modeling (Quick, Dynamic, Multiple, Prune, Exhaustive Prune and RBFN) applied on important features, and the best method with high accuracy selected. The model applied on some meso- and thermophilic xylanase proteins with known temperatures. The results showed 53 features (p value 0.993 to 1.0) such as frequency of glutamine and frequency of hydrophilic residues were important in thermal stability of proteins, and this model can be applied to classify other proteins as either mesophilic or thermophilic with more than 0.95% confidence. The findings show a suitable tool to classify and predict enzymes thermal stability using automated expert systems.

Keywords: Bioinformatics, Feature Selection, Modelling, Neural Networks, Thermostability

1. Introduction

The importance of finding or making thermostable enzymes in different industries has been highlighted. Therefore, it is inevitable to understand the features involving in enzymes' thermostability. Different approaches have been employed to extract or manufacture thermostable enzymes. One of the most important tasks in protein engineering is to understand the important factors for the extreme stability of thermophilic proteins and discriminating them from mesophilic ones [1]. Several methods have been proposed for predicting the stability of proteins upon amino acid substitutions [2]. These methods are mainly based on distance and torsion potentials [3], multiple regression techniques, energy functions [4], contact potentials [5], neural networks [6], support vector machines, SVMs [7], average assignment [3], classification and regression tools [8], backbone flexibility [9] etc. Several attempts have been made to understand the factors influencing the stability from amino acid sequence. It has been reported that increase in number of salt bridges and side chain-side chain interactions [10], counterbalance between packing and solubility [11], aromatic clusters [12], contacts between the residues of hydrogen bond forming capability [1], ion pairs [13], cation-π interactions [14], non-canonical interactions [15], electrostatic interactions of charged residues and the dielectric response [16], amino acid coupling patterns [17], main-chain hydrophobic free energy, and hydrophobic residues [18], in thermophilic proteins enhanced the stability. It has been found that the proteomes of thermophilic proteins are enriched in hydrophobic and charged amino acids at the expense of

polar ones [10]. In spite of these studies, it is necessary to build a system, which derives stability rules for any input data and convert them into prediction.

In recent years, xylanases have received attractable research interest due to their significant application in various industrial processes, such as food, feed, waste treatment, fuel and chemical production, paper and pulp industries [19]. For environmental reasons, xylanases are desirable in that they reduce the amount of chlorine and chlorine dioxide used for bleaching paper pulp. During the bleaching of Kraft paper pulp, the lignin in wood chips is removed by sequential treatments with chlorine, chlorine dioxide, and NaOH. The chlorine and chlorine dioxide create persistent organic chemicals that are toxic to organisms in the waterways close to paper plants and may present health risks to humans as well. Pre-treating paper pulp with xylanases can enhance the efficiency of the chemical extraction of lignin and so reduce the amounts of chlorine and chlorine dioxide required [20]. However, such applications require xylanase(s) with particular properties, the bio-bleaching of paper pulp requires a xylanase with better thermostability [21].

To have a better understanding of features contributing to the enzyme thermal stability, it is necessary to find out the main features responsible for this valuable characteristic. Here we trained different neural networks on few thousands mesophilic and thermophilic enzymes to determine the most important features responsible for thermostability and to find a suitable tool to predict other enzymes' optimum active temperature.

2. Materials and Methods

From UniProt Knowledgebase (Swiss-Prot and TrEMBL) and NCBI databases, amino acid sequences of 4636 proteins - 3069 mesophilic and 1294 thermophilic proteins – were downloaded. 74 features of each protein such as length, weight (Kda), Isoelectric point, aliphatic index, N-terminal amino acid, Half-lives, non-reduced cysteines extinction coefficient and absorption, reduced cysteines extinction coefficient and absorption, counts and frequencies of different atoms, hydrophobic, hydrophilic, other residues, negatively, positively and other charged residues and counts and frequencies of all amino acids were extracted.

Dataset of protein features imported to neural network software (Clementin_NLV-11.1.0.95; Integral solution Ltd), unwanted features filtered, proteins classified into two groups (F = mesophilic and T = thermophilic) according to their optimal temperatures, and this variable set as output and the others as input variables. Feature selection algorithm applied on features datasets; the algorithm considered one attribute at a time to see how well each predictor alone predicted the target variable. The important value of each variable was then calculated as (1-p) where p was the p value of the appropriate test of association between the candidate predictor and the target variable. The association test for categorized output variables was different from the test for continuous ones. When the target value was categorical, p values based on the F statistic were used. The idea was to perform a one-way ANOVA F test for each predictor; otherwise, the p value was based on the asymptotic t distribution of a transformation on the Pearson correlation coefficient. The predictors were then labelled as 'important', 'marginal', and 'unimportant' with values above 0.95, between 0.95 and 0.90, and below 0.90, respectively.

The neural networks used here were feed-forward neural networks, also known as multilayer perceptions. The neurons in such networks were arranged in layers, typically, one layer for input neurons (the input layer), one or more layers of internal processing units (the hidden layers), and one layer for output neurons (the output layer). Each layer was fully interconnected to the preceding layer and the following layer. Back propagation of error, based on the generalized delta rule [22] was used to train the models. For each record presented to the network during training, information (in the form of input fields) was fed forward through the network to generate a prediction from the output layer. This prediction was then compared to the recorded output value for the training record, and the difference between the predicted and actual output was propagated backward through the network to adjust the connection weights to improve the prediction for similar patterns. Seeking for the network architecture with less complexity and better accuracy, we used six different algorithms to generate our models. They were Quick method - a single neural network with one hidden layer containing max (3,(ni + n0)/20) neurons –, Dynamic method - the topology of the network changed during training, with neurons added to improve performance until the network achieved the desired accuracy –, Multiple method - the networks were trained in pseudoparallel fashion after initialization –,

Prune method - conceptually, the opposite of the dynamic method –, Exhaustive Prune method - a special case of the prune method, with two hidden layers and Radial Basis Function Network (RBFN) method - a special kind of neural network with the hidden or receptor layer consists of neurons that represent clusters of input patterns on radial basis functions.

In the modelling phase, our dataset of proteins was partitioned into training and testing groups (90% to 10% ratio) on a random basis. To evaluate the validity of neural networks, in half of them, a validation partition was also created; in these networks three partitions were used (training, testing and validation partitions, 80%, 10% and 10%, respectively). All of the above mentioned methods were then applied, and the results were compared. To study the possible impacts of data preparation on neural network modelling of the proteins' thermostability, this process was repeated on datasets without feature selection as well.

As a result, 24 models were created by 6 methods on 4 different datasets (dataset with no modifications, dataset with important features only, dataset with validation set, and dataset with important features and validation set), and the simplest and the most accurate model was chosen. Finally this model applied on dataset of 80 xylanase enzymes whose optimum temperatures were cited in literatures to evaluate the accuracy of the model in prediction of optimum temperature.

3. Results

The average length of proteins studied here was 330.7±223.05 amino acids with maximum and minimum length of 1983 (GLTS_SYNY3) and 14 (LPF2_ECOLI and LPW_ECOLI) amino acids, respectively. The average weight of proteins was 37±24.6 Kda with average isoelectric point of 7.8±1.86 and aliphatic index of 96.93±15.99. In 98.45% of proteins the N-terminal amino acid was Methionine (Met), in 0.6% of them the same position occupied by Alanine (Ala) and in 0.35% Serin were the N-terminal amino acid. Average of non-reduced Cysteine (Cys) extinction coefficient, non-reduced Cys absorption, reduced Cys extinction coefficient and reduced Cys absorption features were 54.95±1.19, 0.89±0.02, 41.05±0.89 and 0.89±0.02 (Mean±SD), respectively.

The average count of sulphur, carbon, nitrogen, oxygen and hydrogen were 10.92±0.24, 196.05±4.23, 369.69±7.97, 383.86±8.30 and 90.64±1.96, while the average count of hydrophobic, hydrophilic and other residues were 170.11±3.63, 72.72±1.57, 87.54±1.89 (Mean±SD), respectively. The highest average count of amino acids belonged to Leucine (Leu) (33.48±0.72) and the lowest average accounted found to be Cys (3.05±0.07).

The results showed in Quick method a network with 52 neurons in input layer, 1 neurons in hidden layer one and 1 neuron in output layer was generated with 89.53% estimated accuracy and frequency of Glutamine (Gln) (p value of 0.58) as the most and the frequency of hydrophobic residues (p value of 0.01) as the least important input features. In Dynamic method a network with 52 neurons in input layer, 8 and 5 neurons in hidden layer one and two, respectively and 1 neuron in output layer generated with high accuracy (91.35%) and frequency of Arg (p value of 0.58) and the count of Ala (p value 0.027) as the most and the least important features. A network with 52, 28 and 1 neurons in input, hidden layer one and output layers generated in Multiple method with estimated accuracy of 90.99% and the frequency of Arg and the count of sulphur (p value of 0.60 and 0.025) as the most and the least important features. In Prune method a complicated network created with 37, 30 and 1 neurons in input, hidden one and output layers with 91.035% accuracy and the frequency of Arg (p value 0.64) and the count of sulphur (p value 0.041) as high and less important features. In RBFN both network complexity and accuracy were lower than Prune method (52, 1, 1 neurons, respectively and 86.93% accuracy) with the frequency of Glu and the frequency of Leu (p value 0.445 and 0.154) as both ends features. Finally in Exhaustive Prune method the network had 52, 1, 2 and 1 neurons in output, hidden layer one, two and output layers with 90.723% estimated accuracy and the frequency of Arg (p value 0.64) and the count of Phe (p value 0.019) as the most and the least important features.

The results of feature selection showed that 52 features were categorized as important (value more than 0.99) and just one attribute (the count of hydrophilic residues, value 0.91) defined to be marginal and the rest unimportant. The important features were frequency of Gln, hydrophobic residues, Glu, other hydrophobic

residues, Lys, Val, Arg, Asp, Ala, His, Leu, Tyr, Ser, positively and negatively charged residues, Trp, Phe, Sulphur, Met, Ilu, Gly, hydrophobic residues, Pro and Cys and the count of Glu, positively and negatively charged residues, Lys, His, hydrophilic residues, Thr, Arg, other charged residues, Ala, Asp, Val, Tyr, Trp, reduced and non-reduced cysteines absorption, Leu, Ser, sulphur, Cys, Met, Phe, Ilu Half-life of mammals while the count of hydrophobic residues classified as marginal feature.

Considering the analytical and performance evaluation of different models examined here, we found Dynamic model generated with feature selection and validation set as a good tool to check other proteins' optimum temperatures. This model applied on dataset of 110 xylanase enzymes (with known optimum temperature) to test the model performance and the results showed maximum and minimum confidence levels of 0.999 and 0.086 and good average confidence level of 0.952±0.181 (Mean±SD) in defining new proteins' optimum temperature.

4. Discussion

Thermostable enzymes are gaining wide industrial and biotechnical interest due to the fact that they are more stable and thus generally better suited for harsh process conditions [23]. The concept of thermostability is, however, not very clear, and the thermostability is a relative term. The enzymatic activity is known to increase with increasing temperature up to the temperature at which inactivation starts to occur [24]. Thermostability is usually defined as the retention of activity after heating at a chosen temperature for a prolonged period. Thermostable enzymes are produced both by thermophilic and mesophilic organisms. Although thermophilic microorganisms are a potential source for thermostable enzymes, the majority of industrial thermostable enzymes originate from mesophilic organisms [10]. Thermostable enzymes in the hydrolysis of xylan materials have several potential advantages: higher specific activity (decreasing the amount of enzyme needed), higher stability (allowing prolonged hydrolysis times) and increased flexibility for the process configurations [25]. The first two characteristics would expectedly improve the overall performance of the enzymatic hydrolysis even at the range of conventional enzymes active at around 50°C. Thus, carrying out the hydrolysis at higher temperature would ultimately lead to improved performance, i.e. decreased enzyme dosage and reduced hydrolysis time and, thus, potentially decreased hydrolysis costs [26]. Thermostable enzymes would expectedly also allow hydrolysis at higher consistency due to lower viscosity at elevated temperatures and allow more flexibility in the process configurations. Many methods so far have been applied to genetically modify the coding genes to make thermostable xylanase enzymes and the need to define the most active features in making thermostable xylanases has been elaborated [27].

Here we applied different modelling techniques on more than seventy features of mesophilic and thermophilic proteins to find out the most important features contributing to thermal stability. We used different data preparation and various neural networks to find out the relation between proteins' properties and optimum temperature activity.

As mentioned before, various neural networks generated by different data preparation. The most complicated network generated was Multiple method with nearly 81 neurons in network layers; while the simplest networks generated was in Quick method with 55 neurones in network layers. Estimated accuracy of neural networks generated and studied in this study varied between 89% and 91% and the best estimated accuracy (91.35%) gained in Dynamic method, so this method was chosen to test xylanase dataset.

In 40% of neural networks generated here, the frequency of Arg was the most important feature contributing to optimum protein temperature which shows the critical position of this amino acid in thermostability character of proteins.

Regarding the importance of thermostable enzymes in industries and finding suitable tools to suggest, predict and make them, this study showed bioinformatics modelling can be considered as a suitable method. The Dynamic model can be a good candidate to find optimum temperatures of different proteins.

5. Acknowledgements

The authors greatly acknowledge the support of Green Research Center, Qom University, IRAN.

6. References

[1] J. Kongsted, U. Ryde, J. Wydra and J. H. Jensen. Prediction and rationalization of the ph dependence of the activity and stability of family 11 xylanases. *Biochemistry*. 2007, 46 (47): 13581-13592.

[2] M. M. Gromiha, M. Oobatake, H. Kono, H. Uedaira and A. Sarai. Importance of mutant position in ramachandran plot for predicting protein stability of surface mutations. *Biopolymers*. 2002, 64 (4): 210-220.

[3] M. M. Gromiha. Prediction of protein stability upon point mutations. *Biochem Soc Trans*. 2007, 35 (Pt 6): 1569-1573.

[4] T. Hamelryck. Probabilistic models and machine learning in structural bioinformatics. Statistical methods in medical research. 2009.

[5] A. M. Lisewski. Random amino acid mutations and protein misfolding lead to shannon limit in sequence-structure communication. *PLoS ONE*. 2008, 3 (9): e3110.

[6] H. Hayashi, T. Abe, M. Sakamoto, H. Ohara, T. Ikemura, K. Sakka and Y. Benno. Direct cloning of genes encoding novel xylanases from the human gut. *Can J Microbiol*. 2005, 51 (3): 251-259.

[7] A. Garg and G. P. Raghava. A machine learning based method for the prediction of secretory proteins using amino acid composition, their order and similarity-search. *In silico biology*. 2008, 8 (2): 129-140.

[8] L. T. Huang, K. Saraboji, S. Y. Ho, S. F. Hwang, M. N. Ponnuswamy and M. M. Gromiha. Prediction of protein mutant stability using classification and regression tool. *Biophys Chem*. 2007, 125 (2-3): 462-470.

[9] I. W. Davis and D. Baker. Rosettaligand docking with full ligand and receptor flexibility. *J Mol Biol*. 2009, 385 (2): 381-392.

[10] H. M. Yang, B. Yao and Y. L. Fan. Recent advances in structures and relative enzyme properties of xylanase. *Sheng Wu Gong Cheng Xue Bao*. 2005, 21 (1): 6-11.

[11] M. M. Gromiha, M. Oobatake and A. Sarai. Important amino acid properties for enhanced thermostability from mesophilic to thermophilic proteins. *Biophys Chem.* 1999, 82 (1): 51-67.

[12] L. Cicortas Gunnarsson, C. Montanier, R. B. Tunnicliffe, M. P. Williamson, H. J. Gilbert, E. Nordberg Karlsson and M. Ohlin. Novel xylan-binding properties of an engineered family 4 carbohydrate-binding module. *Biochem J.* 2007, 406 (2): 209-214.

[13] Ihsanawati, T. Kumasaka, T. Kaneko, C. Morokuma, R. Yatsunami, T. Sato, S. Nakamura and N. Tanaka. Structural basis of the substrate subsite and the highly thermal stability of xylanase 10b from thermotoga maritima msb8. *Proteins.* 2005, 61 (4): 999-1009.

[14] E. Kiarie, C. M. Nyachoti, B. A. Slominski and G. Blank. Growth performance, gastrointestinal microbial activity, and nutrient digestibility in early-weaned pigs fed diets containing flaxseed and carbohydrase enzyme. *J Anim Sci.* 2007, 85 (11): 2982-2993.

[15] S. Chakkaravarthi, M. M. Babu, M. M. Gromiha, G. Jayaraman and R. Sethumadhavan. Exploring the environmental preference of weak interactions in (alpha/beta)8 barrel proteins. *Proteins*. 2006, 65 (1): 75-86.

[16] F. De Lemos Esteves, T. Gouders, J. Lamotte-Brasseur, S. Rigali and J. M. Frere. Improving the alkalophilic performances of the xyl1 xylanase from streptomyces sp. S38: Structural comparison and mutational analysis. *Protein Sci*. 2005, 14 (2): 292-302.

[17] M. Schubert, D. K. Poon, J. Wicki, C. A. Tarling, E. M. Kwan, J. E. Nielsen, S. G. Withers and L. P. McIntosh. Probing electrostatic interactions along the reaction pathway of a glycoside hydrolase: Histidine characterization by nmr spectroscopy. *Biochemistry*. 2007, 46 (25): 7383-7395.

[18] K. Miyazaki, M. Takenouchi, H. Kondo, N. Noro, M. Suzuki and S. Tsuda. Thermal stabilization of bacillus subtilis family-11 xylanase by directed evolution. *J Biol Chem*. 2006, 281 (15): 10236-10242.

[19] D. Chantasingh, K. Pootanakit, V. Champreda, P. Kanokratana and L. Eurwilaichitr. Cloning, expression, and characterization of a xylanase 10 from aspergillus terreus (bcc129) in pichia pastoris. *Protein Expr Purif*. 2006, 46 (1): 143-149.

[20] B. Sudha, H. Veeramani and S. Sumathi. Bleaching of bagasse pulp with enzyme pre-treatment. *Water Sci Technol*. 2003, 47 (10): 163-168.

[21] G. Ompraba, D. Velmurugan, P. Arumugam, V. Govindasamy and P. T. Kalaichelvan. Homology model of a novel thermostable xylanase from bacillus subtilis-ak1. *J Biomol Struct Dyn*. 2007, 25 (3): 311-320.

[22] J. L. McClelland and D. E. Rumelhart, Parallel distributed processing, vol. 2: *Psychological and biological models, The MIT Press* 1986.

[23] W. W. Wakarchuk, W. L. Sung, R. L. Campbell, A. Cunningham, D. C. Watson and M. Yaguchi. Thermostabilization of the bacillus circulans xylanase by the introduction of disulfide bonds. *Protein Eng*. 1994, 7 (11): 1379-1386.

[24] M. Paloheimo, A. Mantyla, J. Kallio, T. Puranen and P. Suominen. Increased production of xylanase by expression of a truncated version of the xyn11a gene from nonomuraea flexuosa in trichoderma reesei. *Appl Environ Microbiol*. 2007, 73 (10): 3215-3224.

[25] S. Kundu and D. Roy. Comparative structural studies of psychrophilic and mesophilic protein homologues by molecular dynamics simulation. *Journal of molecular graphics & modelling*. 2009.

[26] A. Teplitsky, A. Mechaly, V. Stojanoff, G. Sainz, G. Golan, H. Feinberg, R. Gilboa, V. Reiland, G. Zolotnitsky, D. Shallom, A. Thompson, Y. Shoham and G. Shoham. Structure determination of the extracellular xylanase from geobacillus stearothermophilus by selenomethionyl mad phasing. *Acta Crystallogr D Biol Crystallogr*. 2004, 60 (Pt 5): 836-848.

[27] R. Ruller, L. Deliberto, T. L. Ferreira and R. J. Ward. Thermostable variants of the recombinant xylanase a from bacillus subtilis produced by directed evolution show reduced heat capacity changes. *Proteins.* 2007.