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

Bioinformatics and protein engineering; presenting a few applications employed in our labs

Mansour Ebrahimi [1], Esmaeil Ebrahimie [2], Ahmad Tahmasebi [2], Narjes Rahpeyma Sarvestani[2], Tahereh Deihimi[3]

[1] Bioinformatics Research Group, Green Research Center, Qom University, QOM, IRAN
[2] Department of Crop Production & Plant Breeding, College of Agriculture, Shiraz University, Shiraz, IRAN
[3] Plant Biotechnology Center, College of Agriculture, Shiraz University, Shiraz, IRAN

Abstract. Bioinformatics uses various algorithms and methods to explorer huge amount of biological data in order to help us to understand biological mechanisms. In our labs research groups use bioinformatics tools to investigate and understand why some biological processes are working and what are the most important features contributing to their functions. Of special importance to our researchers are some enzymes and proteins responsible for salinity and drought stresses and thermostability. Different approaches have been employed but they can be classified as follows: a. statistical analyses to understand the significant differences among normal and desired proteins (halophilic or thermostable), b. feature selection algorithms to define the most important features contributing to desired protein activities, c. neural network modelings and tools to train and test different networks in order to correlate between features and protein characteristics and use these networks to predict desired abilities. The results of some research groups have been presented briefly here.

Keywords: Bioinformatics, Neural Networks, Thermostability, salinity stress, drought stress

1. Introduction

Bioinformatics is the emerging field of science growing from the application of mathematics, statistics, and information technology, including computers and the theory surrounding them, to the study and analysis of very large biological, and particularly genetic, data sets [1]. The field has been fuelled by the increase in DNA and protein data generation leading to the massive data sets already generated, and yet to be generated, in particular the data from the human and other genome projects. It has been shown this field can play a critical role in summarizing huge biological data available in databanks and extract applicable models to understand genetic patterns and produce new genes and proteins products based on the modelling results . Proteins are playing critical roles in biological reactions and control all cells activities in living organisms and understanding their functions pave the road of controlling biological functions and producing new products to facilitate favourite reactions. Different approaches have been employed in our labs to understand patterns of activities in some proteins through bioinformatics and statistical tools and summaries have been presented in this paper.

Halophilic bacteria of the genus Halobacterium belong to a separate bacterial kingdom, recently named archaebacteria [2]. The enzymes from extremely halophilic bacteria represent a fascinating example of adaptation. Until now only a few intracellular halophilic proteins have been isolated in a pure state and characterized [3]. Considerable attention has been paid to halophilic enzymes in order to understand how they are stable and active in the presence of very high salt concentrations [4]. A statistical analysis of halophilic proteins and their nonhalophilic homologues has shown an increase of acidic amino acids is a general property of halophilic protein. Other changes such as an increase in the number of small side-chain hydrophobic amino acids, and a decrease in the number of aliphatic side-chain amino acids are also observed [5].

Biotic and abiotic stresses are known to act as a catalyst in producing free radical reactions resulting in oxidative stress in various organisms where reactive oxygen species are produced. In this context, aerobic organisms have developed several non-enzymatic and enzymatic systems to neutralize these compounds. The enzymatic systems include a set of gene products such as superoxide dismutases, catalases, ascorbate peroxidases, glutathione peroxidases, glutathione reductases [6]. Glutathione reductase (GR; EC 1.6.4.2) is widely distributed both in eukaryotes and prokaryotes and it catalyzes the reduction of oxidized glutathione disulfide (GSSG) to reduced glutathione (GSH) using the NADPH as an electron donor [7]. Our general objectives for this study are (i) to investigate the phylogenetic relationships of glutathione reductase through the comparison of genes and protein sequences and structures to the green plants and (ii) to determine the effects subcellular localization on various properties of protein sequence.

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 and 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 [8]. 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. 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 [9].

To have a better understanding of features contributing to the halophilicity, thermostablility, metal transporting and transcriptional proteins, it is necessary to find out the main features responsible for these valuable characteristic. So we have tried to extract various primary, secondary, tertiary and even quaternary of those protein features, classify them and by using various bioinformatics tools (such as feature selection, decision tree and neural networks algorithms and modelling) to determine the features responsible for desired cell actions and to create a suitable tool to predict other enzymes activity regarding those characters.

2. Materials and Methods

From UniProt Knowledgebase (Swiss-Prot and TrEMBL) and NCBI databases, amino acid sequences of 50 halophilic and 50 nonhalophilic proteins, meso- or thermophilic enzymes - 3069 mesophilic and 1294 thermophilic proteins -, transcription factors – 11016 ABA dependent and 205 ABA independent - and antioxidant proteins – superoxidase, dismutase, glutathione reductase and peroxidises - extracted and 350 features of each protein calculated. We categorize these features into four categories as follows:

(A) Primary structure: amino acid composition of the protein sequence.

(B) Secondary structure: the amount of helix structure found within the protein and the number of atoms making up the helices.

(C) Tertiary structure: Ion pairs, hydrogen bonds, disulfide bonds and accessible surface area (ASA).

(D) Extended properties: Normalize the value obtained from (A), (B) and (C) above and the ratio of various other features such as polar/nonpolar.

To find the most important features contributing to the desired protein character (thermostability, halophilicity, transcription factor and antioxidant activity) two separate approaches employed:

1. **Statistical analyses:** Various statistical methods such as descriptive statistics, compare means (T-tests, ANOVA, …), general linear models (univariate, multivariate), correlation, regression, step wise regression and empirical bays used to interpret the relation between protein features and desired characters.

2. **Modelling and Neural network analyses:** dataset of protein features imported to neural network softwares (Clementin_NLV-11.1.0.95; Integral solution Ltd) and Matlab (R2008b, the Mathworks Inc) and the proteins features classified into two groups (F = undesired and T = desired) according to the favourite protein character (thermostability, halophilicity, transcription factor and antioxidant activity), and these variables set as output and the others as input variables. The following modelling applied to these datasets:

- Decision tree: Decision trees work by recursively partitioning the data based on input field values. The data partitions are called branches. The initial branch (sometimes called the root) encompasses all data records. The root is split into subsets, or child branches, based on the value of a particular input field. Each child branch can be further split into sub-branches, which can in turn be split again, and so on. At the lowest level of the tree are branches that have no more splits. Such branches are known as terminal branches (or leaves). The rule browser shows the input values that define each partition or branch and a summary of output field values for the records in that split. Various decision tree algorithms such as C5.0, C&R Tree, Quest and CHAID applied on protein features' datasets.
- Feature selection: Feature selction 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.
- Neural Networks: The neural networks used here were feed-forward neural networks, also known as multilayer perceptrons. Back propagation of error, based on the generalized delta rule [10] 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. 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. As a result, 96 models (24 neural networks for each protein studied here) 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 to evaluate the network accuracy and precision. Neural network models were repeated in both Clementine and Matlab softwares to compare neural network topologies and accuracy.

3. Results

It is not possible to show all results we had so far in detail here so only briefly some parts have been presented.

Results on GR: A detailed comparison of GR genes across plant species revealed a high degree of conservation in gene structure for GR isoforms localized in the same subcellular compartment. The cytosolic GR genes show more comprehensive structures in the genome than its chloroplastic counterparts, that is, the cytosolic GR genes are composed of 17 exons interrupted by 16 introns in the transcribed region but chloroplastic GR genes are composed of 10/9 and 11/10 exons/introns. Transit peptide sequences located in the N-terminus of these proteins facilitate the transfer from the cytosol where they are synthesized, back to the chloroplast organelle. We investigated the targeting signal of chloroplastic isoforms (data not shown) that shown this region is rich in Ser but relatively low in Arg. The alignment of all cytosolic and chloroplastic GRs protein sequences revealed that the general features of GRs, including the residues important in binding GSSG, the redox-active disulphide bridge domain and the conserved arginine residues required for NADP binding, are majority of present. The phylogenetic tree structure clearly revealed that these GRs split into tow main clusters. The first one encompasses chloroplastic isoforms. The second group contains cytosolic isoforms. In addition, we generated a phylogenetic tree of nucleotide sequence that confirmed above result. This suggests that an initial duplication event generated the ancestral genes encoding the chloroplastic and the cytosolic isoforms.

Results on thermostability: 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 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. 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.

Results on Halophilicity: Results showed that a significant differences ($P<0.05$) between ABA dependent and independent enzymes is present and a similar patterns observed between these enzymes and other plants' proteases. Results also revealed a conserve sequence can be aligned in all halophilic proteins.

4. Discussion

Here we employed different modelling techniques and statistical methods to retrieve patterns of activity in various classes of enzymes and proteins. In general we are tying to use these methods to determine the most important features contributing to desirable protein characters such as halophilicity, thermostability and drought stress. The results so far confirmed that these methods can be applied to engineering new proteins and to understand how proteins can contribute to these characters.

5. Acknowledgements

The authors greatly appreciate and acknowledge the support of Bioinformatics Research Group, Green Research Center, Qom University and School of Agriculture, Shiraz University, for supporting the project.

6. References

[1] J. Li, Y. Jiang, R. Fan. Recognition of Biological Signal Mixed Based on Wavelet Analysis. In: Y. Jiang, et al (eds.). *Proc. of UK-China Sports Engineering Workshop*. Liverpool: World Academic Union (World Academic Press). 2007, pp. 1-8. (Use “References” Style)

[2] R. Dewri, and N. Chakraborti. Simulating recrystallization through cellular automata and genetic algorithms. *Modelling Simul. Mater. Sci. Eng*. 2005, **13** (3): 173-183.

[3] A. Gray. *Modern Differential Geometry*. CRE Press, 1998.