A Neural Networks System for Enzymes Classification

[1]Wagner Rodrigo Weinert, [2]Heitor Silvério Lopes
[1,2] Laboratório de Bioinformática / CPGEI
Centro Federal de Educação Tecnológica do Paraná (CEFET-PR)
Av. 7 de setembro, 3165 – 80230-901 Curitiba (PR), Brazil
[1] weinert@cpgei.cefetpr.br, [2] hslopes@cpgei.cefetpr.br

ABSTRACT

This paper describes a biomolecular classification methodology based on Multilayer Perceptron (MLP) Neural Networks [1]. In particular, this paper describes the application of the system to the classification of enzymes into six families. Enzymes are a subclass of proteins that are specialized in catalytic activity [2]. The primary goal of classification, in this case, is to infer the function of an (unknown) enzyme by means of the analysis of its structural similarity to a given family of enzymes. In general, proteins can have similar structures one each other due to their common evolutionary origin [3], but this does not mean that they have similar functions. Also, proteins of the same family seldom have variable lengths of their primary structure (from tenths to thousands of amino acids). These facts, among others, make the problem of protein classification a real challenge for which there is no efficient computational solution.

In this paper, an artificial neural networks system with MLPs is used for classification. First, a new codification scheme was devised to convert a string of amino acids (the primary structure of the enzyme) into a real-valued vector. This vector is directly presented to the first layer of the neural network. This is an improvement over other methods that first try to extract relevant information from the sequence, such as similarity, before using it for classification.

The codification procedure basically consists in the determination of a numerical alphabet of real values in the range [0..1] (excluding zero), representing the 20 amino acids. This was accomplished using the hydrophobicity scale of Kyte-Doolittle [4].

The neural system consists of a group of fixed-topology MLPs. Each MLP had 40-81-6 neurons respectively in the input, hidden and output layer. The number of MLPs was set for the problem according to the length of the longest sequence to be classified in each experiment. After being encoded, enzymes are juxtaposed presenting the MLPs. But, since they have different lengths, a policy of weights was established for computing the final classification based on partial outputs of each neural network.

To evaluate the performance of the neural system, some experiments were done and results were compared with Hidden Markov Models (HMMs) using the software HMMER 2.2 [5]. A total of 8339 enzymes (600 for training and 7739 for validation), classified into 6 families, were used in the experiments. All data were obtained from the Protein Data Bank (PDB), release 102 (Oct.2002). A random sample of 100 enzymes was drawn from PDB for each family, to form the training set. All remaining enzymes of each family constituted the validation set (the number was different for each family). A randomized 5-fold cross-validation procedure was done using the same data for both methods, and the performance was measured using sensitivity (Se) and specificity (Sp). Table 1 shows these results. Each cell represents the average result (over the 5 experiments) followed by the standard deviation.

Table 1: Comparative results of classification.

	HMM		**Neural Networks**	
Enzyme class	**Se**	**Sp**	**Se**	**Sp**
Oxidoreductases	40,5±33,2	73,6±23,5	82,0±5,8	91,2±1,1
Transferases	23,5±13,9	93,1±5,3	72,1±4,6	92,5±3,3
Hydrolases	51,3±13,5	82,6±12,2	73,3±5,1	95,4±1,9
Lyases	38,4±14,8	95,2±4,3	73,8±7,1	95,7±0,9
Isomerases	58,7±8,8	94,7±5,3	76,2±6,7	96,9±0,8
Ligases	54,9±10,8	88,3±15,9	70,5±8,8	97,2±0,6
average	44,5±20,1	87,9±14,3	74,6±7,0	94,8±2,7

Results demonstrate that neural networks are suited for the biomolecular classification task and that they display a much better performance than HMMs. The high values of standard deviation found in all experiments with HMMs indicate its strong dependence on the training set. The same was not observed for the neural system suggesting its robustness.

REFERENCES

[1] Fausett, L. **Fundamentals of Neural Networks.** Upper Saddle River: Prentice-Hall, Inc., 1994.

[2] Lehninger, A. L., Nelson, D. L., Cox, M. M. **Principles of Biochemistry with an Extended Discussion of Oxigen – Binding Proteins**. 2. ed. New York: W.H. Freeman & Co., 1993.

[3] Murzin, A. G., Brenner, S. E., Hubbard, T., Chothia, C. A structural classification of proteins database for the investigation of sequences and structures. **Journal Molecular Biology**, vol. 247, pp. 536-540, 1995.

[4] Kyte, J., Doolittle, R. A simple method for displaying the hydropathic character of proteins. **Journal of Molecular Biology**, vol. 157, pp. 105-132, 1982.

[5] Eddy, S. R. Profile hidden Markov models. **Bioinformatics**, vol. 14, pp. 755-763, 1998.