Improvements in Determining Protein Subcellular Location

Eduardo Battistella[1], Adelmo Luis Cechin[2]
[1,2]Programa Interdisciplinar de Pós-Graduação em Computação Aplicada – PIPCA – Centro de Ciências Exatas e Tecnológicas – Universidade do Vale do Rio dos Sinos (UNISINOS) – Av. Unisinos, 950 – 93.022-000 – São Leopoldo, RS – Brasil
{eduardob,cechin}@exatas.unisinos.br

ABSTRACT

Knowing where a protein occurs in the cell is an important step towards understanding its function [1]. Hence, a method for accurately predicting the subcellular location would be valuable in interpreting the data being provided by sequencing projects. Among some methods (e.g. search for signal peptides, infer the location by sequence homology) the correlation between the total aminoacid composition of proteins and its subcellular location is the most studied one. In this context, this paper introduces a new physical-chemical attribute relevant to the process of determining the protein subcellular location when utilizing the yeast database. This was achieved by developing a series of tests involving multiples Artificial Neural Networks (ANN) and the subsequent use of Linear Discriminant Analysis (LDA) as a way to explain the results reported by the ANNs. Two improvements were obtained: first, better classification scores than those previously produced by other works; second, a better choice of attributes resulting in a further improvement.

The yeast database [4] was used by two previous works [2][3] where both tried to predict the protein subcellular location involving different techniques. There is some doubt concerning to the methodology used during the test of the ANNs developed by Cairns [2]. The approach developed by this paper utilized an adequated training process involving n-fold-cross-validation where "n" is suitable to the data available. Instead of using one network, multiple networks were developed. By doing this, the impact of the interferences inter-classes during the learning process was minimized. Additionally a series of tests was done to explore the space of possible network architectures.

A misclassification was observed between some classes. For example, the sites Cytoplasm and Nuclear interfered resulting in a lot of classification errors. At this point LDA was used in an attempt to explain these errors.

The results obtained by LDA reached the identification of one completely useless attribute in the database and the characterization of conflicts involving the values of the attributes that were supposed to identify the location. Identified the necessity of a new physical-chemical attribute to improve the results we tested 16 new ones obtained from the Yale University database in a subset of 80 randonly choosed proteins containing 40 correct cases (20 cytoplasm and 20 nuclear) and 40 incorrect (20 cytoplasm that were reported as nuclear and vice-versa). The results using the protein´s "Isoeletric Point" reported an improvement from 60% to 72.5% in the classification score.

As future works we intend to remake the tests involving the ANNs by replacing the useless attribute by the Isoeletric Point.

REFERENCES

[1] ANDRADE, Miguel A.. O'DONOGHUE, Seán. ROST, Burkhard. Adaptation of Protein Surfaces to Subcellular Location. Journal of Molecular Biology, 276, 517-525, 1998.

[2] CAIRNS, Paul. et al. A Comparison of Categorisation Algorithms for Predicting the Cellular Localization Sites of Proteins. Proceedings of the 12th International Workshop on Database and Expert Systems Applications, 296-300, 2001.

[3]NAKAI, Kenta. HORTON, Paul. A Probabilistic Classification System for Predicting the Cellular Localization Sites of Proteins. Intelligent Systems in Molecular Biology, 109-115, 1996.

[4] Yeast database. www.ics.uci.edu/~mlearn/MLSummary.html.