Poster K-15

A New Methodology for Regulatory Network Inference

Authors:
Ígor L. Almeida (*PIPCA, Universidade do Vale do Rio dos Sinos - São Leopoldo, Brazil*)
Adelmo L. Cechin (*PIPCA, Universidade do Vale do Rio dos Sinos - São Leopoldo, Brazil*)
Cláudia K. Barcellos (*Hewlett-Packard Labs - Porto Alegre, Brazil*)
Irene S. Gabashvili (*Computational Biosciences Research, Hewlett-Packard Labs - Palo Alto, USA*)

Short Abstract: DNA array technologies are powerful tools to monitor simultaneously the expression patterns of thousands of genes. The challenge is to extract useful information from the vast amounts of data generated. This work proposes a new Markov Chain-based approach for analysis of microarray data.

Long Abstract:
The genomic projects are growing faster and with it the amount of data generated by the experiments. Microarray is one of them. Efficient computational techniques to analyze this data have become really necessary.

The microarrays were developed on the 90's and were able to analyze the gene expression levels of thousand of genes at the same time. Using them, it is possible to study the gene expression patterns that are the base of cellular physiology. This work intends not only to analyse which genes are or not being activated but also the interaction in the form of causal-consequence relations that could exist among them.

The temporal relation among the gene expression levels will be expressed in the form of a Markov Chain extracted from a Recurrent Neural Network (RNN) trained with microarrays data. The network plays the role of a recurrent nonlinear model. Although there are many public database containing time series of gene expression levels, usually they do not have more than 80 time samples, what makes the acquisition of a good model a difficult task. Small time series may not be enough to extract all the relation existing in the data, so it was decided to work with a set of 80 gene expression vectors for 2467 yeast genes that were selected previously by other authors. The Data were generated from spotted arrays using samples collected at various time points during the diauxic shift, the mitotic cell division cycle, sporulation, and temperature and reducing shocks. This data is part of the Stanford Microarray Database (SMD) and are available at http://rana.stanford.edu/clustering.

Microarray data have typically two drawbacks. First, it has a lot of noise. Second, it contains a lot of redundant data with correlated gene expression levels. If this data is used as input parameters of a RNN, then the first step to treat the data is a dimensional and noise reduction, in order to obtain useful results in the process of Markov Chain extraction. To guarantee significant results, a lot of techniques have been proposed with success. Among them, there are Hierarchical Clustering, Nearest Neighbors and Self-organizing Maps (SOM).

Considering this, to find a reduced data set that could express the principal characteristics of the complete data base, the data was processed with Self-Organizing Maps (SOM). The motive to use this tool is to discover the more prominent features in this data and to determine how many they are.

Afterwards, with the Markov Chain representing the temporal behavior of the gene expression levels, the cause relations and influences among genes (or among groups of them) may be discovered. This knowledge promotes a better understanding about the organism under analysis. First, by understanding the network of gene activation, it is possible to predict the reaction of the organism (activated genes) under the influence of drugs, environment conditions or phase of the life cycle. Second, the temporal relation in the form of a Markov Chain allows the scientist to understand and predict under which conditions the organism changes its network of gene expression levels, and therefore how it adapts to adverse conditions. To extract this information without the use of computational techniques is impracticable because the great amount of data.

In the extractions process, it is necessary to train a Recurrent Neural Network (RNN), which is a nonlinear dynamic system and are able to store the dynamics of the gene expression levels in the form of internal weights. The state extraction is related to the division of the work space of the RNN, or to clustering methods. The clustering method investigated is the fuzzy clustering. This state representation is composed of a set of membership functions and piecewise linear approximations. The membership value can be interpreted as a measure of validity of the linear approximation.

Two compromises have been reached in the choice of the number of states. Normally, larger states contain more data and represents larger regions, resulting in a more spare representation, which is also easier to understand but represents the data in a less exact way. Smaller states, on the contrary, represent few data, very exact, but many of them are required to represent the same input space making them more difficult to understand.

The state transitions can be determined using the time series of the gene expression levels. So, to compute the transition probability values, we used the transition frequency. The probabilities are associated to each transition among the states, as well as the transitions that allow the system staying at the same state.

The RNN used for extraction of the Markov Chain consists of three layers (11-5-11) and was based on the Jordan Architecture. For the extraction, three membership functions were defined for each neuron, but only 24 were used to represent the training data. The Markov Chain obtained has 5 states, which represents 95% of sampled data. From this Markov Chain, we may extract the regulatory network corresponding to each state. The first state represents the main regulating relations among the investigated genes and it can be showed as a graph, where it’s possible to observe different influences among the gene groups (nodes) and the influences (arcs) both in the form of activation (positive arcs) or inhibition (negative arcs). There was no direct relation between the 11 clusters (gene groups) and a specific metabolic pathway.