English

Modeling human cancer-related regulatory modules by GA-RNN hybrid algorithms

Jung-Hsien Chiang* and Shih-Yi Chao

Department of Computer Science and Information Engineering, National Cheng Kung University, Taiwan
**Email:jchiang@mail.ncku.edu.tw**

Modeling cancer-related regulatory modules from gene expression profiling of cancer tissues is expected to contribute to our understanding of cancer biology as well as developments of new diagnose and therapies. Several mathematical models have been used to explore the phenomena of transcriptional regulatory mechanisms in Saccharomyces cerevisiae. However, the contemplating on controlling of feed-forward and feedback loops in transcriptional regulatory mechanisms is not resolved adequately in Saccharomyces cerevisiae, nor is in human cancer cells. In this study, we introduce a Genetic Algorithm-Recurrent Neural Network (GA-RNN) hybrid method for finding feed-forward regulated genes when given some transcription factors to construct cancer-related regulatory modules in human cancer microarray data. This hybrid approach focuses on the construction of various kinds of regulatory modules, that is, Recurrent Neural Network has the capability of controlling feed-forward and feedback loops in regulatory modules and Genetic Algorithms provide the ability of global searching of common regulated genes. This approach unravels new feed-forward connections in regulatory models by modified multi-layer RNN architectures. We also validate our approach by demonstrating that the connections in our cancer-related regulatory modules have been most identified and verified by previously-published biological documents.

It is acknowledged that the causes of heterogeneity genetic-related circumstances, such as the cell cycle, or cancer diseases, are products of complex interactions between genes over time. The analysis of cancer-related gene expression data will thus become increasingly widespread. When appraising approaches for discovery of cancer-related regulatory modules, the amount and type of sources of data must be taken into account. Besides, the approach must be capable of handling noisy and high dimensional gene expression data. The approach described here has been shown to be effective with real-world expression data. The stochastic nature of GA means that the same results can not be expected from each run of the algorithm, and the GA is run for a fixed number of generations for each output of regulatory modules. However, to increase the number of genes that the GA can select from, it could require more GA generations. As a result, increasing the GA generations also increases the computational time, although it does show that results on microarray data can be discovered correctly by the GA used in our approach. In addition, this approach builds modules "piece by piece", that is, regulatory module by regulatory module. We discover all the formed units one by one and eventually join these units by their simultaneously existing transcription factors(TF). The above-mentioned contents are the advantages of

generating smaller but more precise regulatory modules, in that each of the paths or the units (or genes) in the modules can be seen without being masked by other connections. It is not the same as traditional complicated regulatory relationships, which are too many to visualize as a network to yield useful information in a digestible format for biologists. Following diagram depicts the framework and flowchart of our approach.
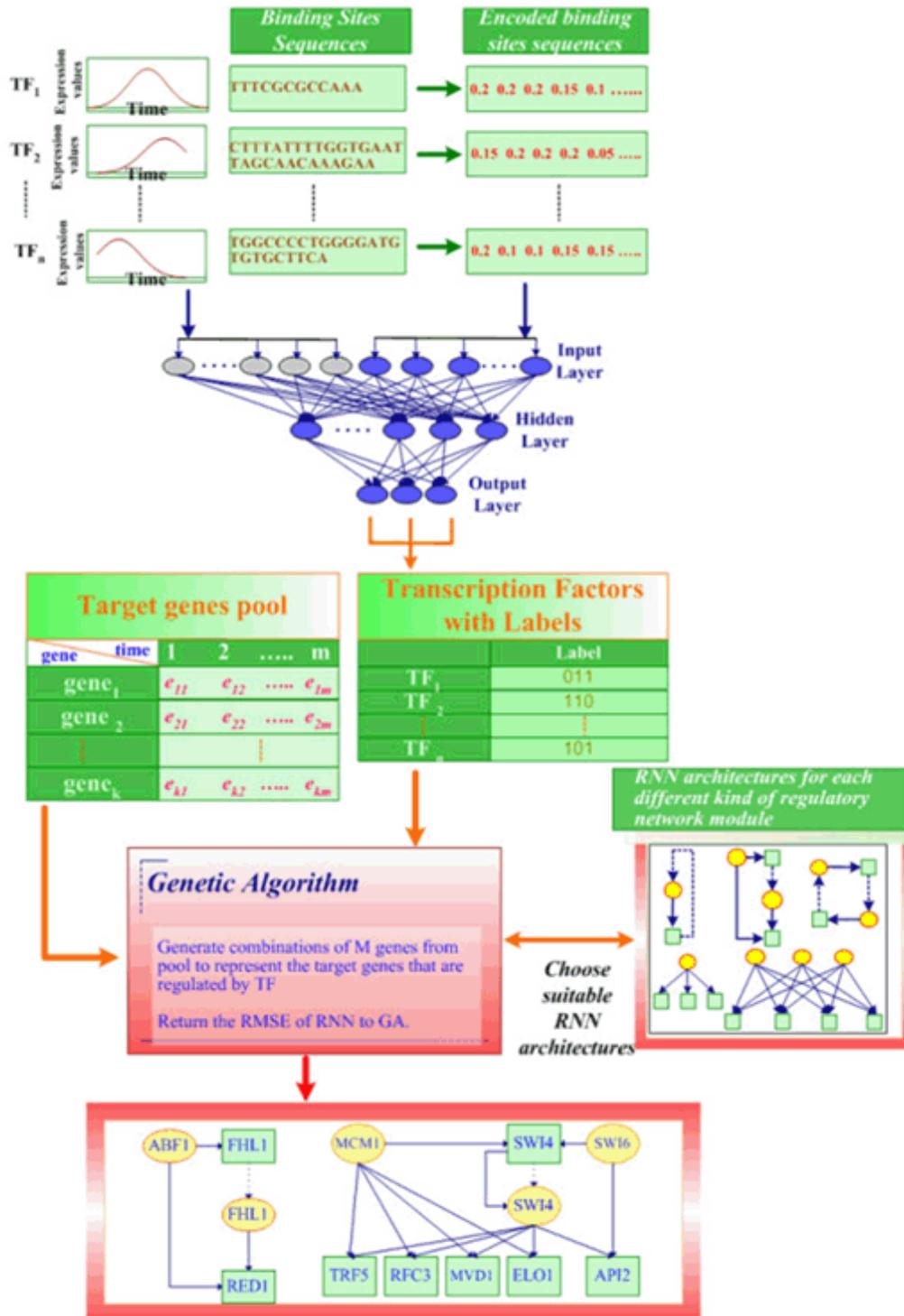
We combine the GA and RNN computing approaches to construct the cancer-related regulatory modules in silico. Upon the microarray data and the sequences of transcription factor binding sites, the approach has been shown to be able to accurately fit the data on which it is trained. We also observe that some TFs play critical roles in various motifs. In other words, some functions of TFs are fit for several kinds of regulatory modules. We then adopt these characteristics by training the radial basis function classifier

for categorizing TFs. Additionally, the experimental results have proven that the GA-RNN hybrid algorithm has the capability of constructing the feedback and feed-forward regulatory modules. RNNs with diversified architectures indicate varied regulatory mechanisms to construct complete regulatory modules with feedback and feed-forward controls. Combining modified RNN with GA, it provides the global searching capacities to find proper target regulated genes for some TFs. The chromosomes that the GA used are combinations of target genes and the crossover and mutation operators used by GA on all chromosomes alter the choice of output gene combinations. This approach is on the basis of both gene expression data and sequences data, so it is time significant and binding region significant data analysis. Summing up, since this method has been previously shown to also classify TFs as well and then construct regulatory modules, it can be considered a candidate multipurpose tool for microarray expression data analysis.