An Integrated Platform to Construct Transcriptional Network from Gene Expression Data

Tao-Wei Huang[1], Hwa-Sheng Chiu[1], Ming-Hong Lin[2], Han-Yu Chuang[1], Chi-Ying F. Huang[2], Cheng-Yan Kao[1]

Keywords: gene expression data, pathway, gene network, transcriptional network

1 Introduction.

In the post-genome era, systems biology is an emergent field for understanding of biological systems at whole system-level instead of single gene or protein. There are numbers of exciting and profound issues that are actively investigated such as the gene regulation and biochemical network. In this study, we focus on the transcriptional networks. The traditional clustering algorithm is applied to the gene expression data to acquire some groups of genes. The signature algorithm [2] can be used to refine and extend these groups of genes generated by clustering algorithm. When a group of genes with common *cis*-regulatory binding motif in promoter region, and we can assume these genes are co-regulated by the same *trans*-regulatory factor. Not only the binding motif information is investigated but also known pathway, protein interactions and cellular component of Gene Ontology information are referred [3]. We performed this idea and approach to find the most upstream regulators. Appling this approach to cancer research, these genes may be novel oncogenes. We integrated some biological databases and also provided some bioinformatics tools as web service for biologists to predict the transcriptional networks *in silico*. By this approach, we found some important transcriptional module and transcription factors, and we will exam it by RNAi experiment *in vitro* or *in vivo*. Besides, the more details of our result can be found on our website (http://insilico.csie.ntu.edu.tw:9999/RECOMB2004/).

2 Method.

In order to obtain more precise transcriptional networks, we integrated the some well-know biological databases locally, updated periodically and developed some bioinformatics software packages as followings:

(1) MotifFinder: This package can accept a list of genes of interest as input. The upstream promoter sequences of these genes are extracted from DBTSS database. The range of promoter sequences takes as parameter from -3000 to +1000. It will also process the retrieved result from TRANSFAC and find the common *cis*-regulatory elements and corresponding *trans*-regulatory factors as output.
(2) GOFinder: The given input is a list of genes of interest. The biological process, cellular component, and molecular function categories of Gene Ontology is shown. Besides, the biologists can search the gene ontology information with different ontology level from 1 to 5.

[1] Dept. of Computer Science and Information Engineering, National Taiwan University, Taipei 106, Taiwan.
E-mail: {d90016, r91031, r90002, cykao}@csie.ntu.edu.tw
[2] Division of Molecular and Genomic Medicine, National Health Research Institutes, Taipei 115, Taiwan.
E-mail: {daycolin, chiying}@nhri.org.tw

(3) PathwayFinder: The pathway package can accept a list of genes of interest as input. All the BioCarta and KEGG pathways containing these genes will be reported in a descendant order according to the number of input genes in a pathway.
(4) ProteinInteractionFinder: The interaction tool can take a list of genes of interest as input. The protein-protein interactions information from DIP and the annotations of these interacting proteins retrieved from SWISS-PROT will be shown.

We use the bottom-up approach to construct the transcriptional networks from small transcriptional modules. The rectangle box stands for a transcriptional module. The gene or protein represented as white ellipse, and the black ellipse is the protein with interactions (Figure 1). The detail processes described as following:

Step1. As a first step for further analysis, we applied the k-means clustering algorithm to gene expression data. The genes in the same cluster *c* are possibly and potentially co-regulated.
Step2. After generating the k clusters, we applied the signature algorithm to each cluster *c* to obtain a new cluster *c*' as 'transcriptional module' initially. The signature algorithm can be used to extend and refine partial knowledge about a pathway module. The genes in the same module *c*', therefore, are more potentially co-regulated.
Step3. Some useful bioinformatics tools are provided to analyze these modules and more annotations are extracted from GeneCards, GO, SWISS-PROT, NCBI for biologists. Genomics information such as binding motif and Gene Ontology are provided. Proteomics information such as pathway and protein-protein interactions is provided. Combining the biological information, the biologists can infer the regulators and construct transcriptional networks more precisely.
Step4. Take the regulating genes or proteins as newer transcriptional module, and apply Step3 to these modules recursively.

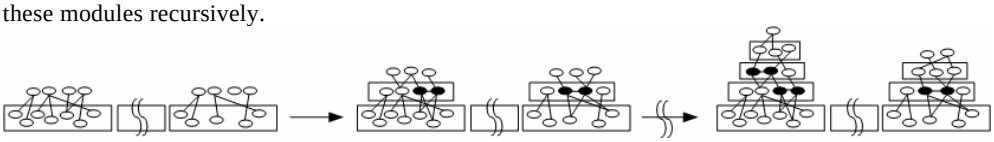

Figure 1: Construct transcriptional network from transcriptional module iteratively.

3 Results.

HCC microarray dataset, obtained from Stanford Microarray Database, contains 1648 differentially expressed genes in HCC vs. nontumor liver samples as analyzed by Chen *et al*. [1]. By our approach, we found some important transcription modules and some regulators can regulate these modules. From further literature review, we also found the regulators involve in hepatocyte regeneration upon partial hepatectomy. Besides, we will exam it by RNAi experiment *in vitro* or *in vivo*. This approach may provide novel targets involved in the carcinogenesis of HCC.

References

[1] Chen X, Cheung ST, So S, Fan ST, Barry C, Higgins J, Lai KM, Ji J, Dudoit S, Ng IO, Van De Rijn M, Botstein D, Brown PO. 2002. Gene expression patterns in human liver cancers., *Molecular Biology of the Cell*, Jun;13(6): pp. 1929-1939

[2] Ihmels J, Friedlander G, Bergmann S, Sarig O, Ziv Y, Barkai N. 2002. Revealing modular organization in the yeast transcriptional network., *Nature Genetics*,Aug;31(4): pp. 370-377

[3] Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N. 2003. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data., *Nature Genetics*, Jun;34(2) pp. 166-176