Research on gene interaction based on graph structure

Toshiaki Ohtani (510020)

School of Information Science,
Japan Advanced Institute of Science and Technology

February 8, 2007

Keywords: gene interaction, microaray, binding site.

A gene indicates a subregion with a specific role in a DNA sequence. By its expression, protein related to some function of organism is synthesized. There are genes that synthesize proteins that promote or regulate expression of themselves or other genes. They are called transcription factor. Such regulation mechanism among genes through transcription factors is called gene interaction mechanism. By identifying the interaction among a large number of genes, we can model various biological reactions in the cell as one concrete system. This may lead to development of new medicine. By this reason, improving accuracy in the estimation of gene interaction will be required.

Existing researches for the estimation of gene interaction mainly uses two kinds of approaches, one is using biological properties, and the other is using mathematical modeling techniques. It is known that a transcription factor recognizes a specific sequence pattern in the gene regulatory region and binds to it. In the first approach, multiple alignment is used for discovering common motifs in regulatory regions of genes that are already known to be co-regulated. Using the motifs, genes having the motifs in their regulatory region are found, and are estimated as putative co-regulated genes. In the second approach, using expression profiles of genes measured by DNA microarray technology, mathematical models representing interaction and regulation mechanism among genes are estimated by computational methods. For the mathematical formalism, Boolean networks,

in which the state of each gene is binalized and dependence between genes is represented by logical functions, and Bayesian networks, in which conditional probabilities between states of genes are represented by a directed acyclic graph, are proposed. However,estimation of Boolean networks requires a large number of time-series expression profiles which are difficult to obtain. Bayesian networks are widely used in this field now. Since it is no more than a model for representing behavior of the expression levels only, there can be many models that can equally explain the same data. Since error ratio of microarray data may reach to 30-40% in the present experiment technology, it is difficult to identify an accurate model using the microarray data only.

The aim of this research is to improve the method proposed Hiraishi and Doi. It is a method for estimating gene interaction using combination of DNA microarray data and sequence analysis on the genome. The procedure of this method is as follows. First, we select a set of genes whose correlation coefficients in the expression profiles compared with a given gene a are high. Next, we compute similarities between all pairs of windows in regulatory regions of gene a and each gene in the selected set of genes, where a window is a subregion of length l. The maximum similarity for a fixed position in gene a is also computed. Then peak positions are selected as those with a high maximum window similarity which is also a local maximal in its neighbor. If there is a subregion that contains peak positions almost everywhere in it, then we estimate it as a binding site. The genes having a window with a high similarity in the site are putative co-regulated genes. The method uses not only statistical information on the correlation in the expression profiles, but also biological knowledge that co-regulated genes have similar sequence patterns in their regulatory regions. In addition, it can estimate binding sites and binding sequences too. However, the following problems are unsolved in the method. 1. It may happen that known binding sites are not selected as a peak position. 2. The final result sometimes contains many candidates for binding sites and we are not able to distinguish the true binding site from others. 3. Various thresholds are used in the method but they are not decided based on sufficient investigations.

For these problems, we first investigate the case that the window includ-

ing the binding site is not selected in peak positions. Then we clarify the relation between window positions and the position that gives the maximum window similarity, and improve the selection rule for peak positions. The peak positions which are not selected in the existing method can now be selected in the new rule. Next, in order to reduce the number of candidate for binding sites, we propose a new approach using window similarities between genes. Concretely, we consider a graph representing the similarities between genes. If there are a set of co-regulated genes, then they probably compose a clique in the graph. Using this fact, we find a clique with more than two nodes in it, and select genes in each clique as putative co-regulated genes. As a result, the number of candidates for binding sites can be reduced to 1/3 comparing with the existing method. In addition, we also propose to use randomly generated sequences to find binding sites, i.e., if the actual similarity of some position is specifically higher than that obtained from the random sequences, then the position is a candidate for binding sites. Using this idea, we can estimate in high accuracy the binding sites having a relatively longer binding sequence. Finally, we optimize threshold values through investigations.