研究計画提案書（修士）

平成 15 年 2 月 25 日

氏名	内田　靖哉	学籍番号	250011
主テーマ指導教員	佐藤　賢二	副テーマ指導教員	亀岡　秋男
主指導教員	小長谷　明彦	副指導教員	佐藤　賢二

＜研究テーマ＞

Interpreting the patterns of gene expression of *Saccharomyces cerevisiae*

＜研究の目的＞

The whole genome sequences of the yeast were released in 1996, but how genes control one another is far from understood. We now have tools, such as microarrays, to monitor the gene expression, but how to reconstruct the regulatory pathways governing observed patterns of expression is still an open question. Therefore, my research objective is to cluster the expressed genes into their unique groups and to explore the regulatory pathways of the yeast *Saccharomyces cerevisiae*.

＜研究の背景・特色＞

In the past decade, there has been an intense effort to comprehensively catalogue the expressed genes in *Saccharomyces cerevisiae* and to determine the absolute and relative abundance of transcript and protein levels under various cellular conditions. Through the emergence of microarray technology, we can now observe the expression levels of thousands to tens of thousands of genes in a single array. However, due to the complexities of the organism, ample amount of information is not being interpreted from this microarray data.

Currently, researchers make two assumptions when interpreting microarray data: (1) There is evidence that many functionally related genes are coexpressed. (2) Coexpressed genes may give information about regulatory mechanisms of the organism. To find these coexpressed genes, two approaches are being used: pattern discovery (unsupervised methods) and class prediction (supervised methods). Both approaches do show some results, but these results do not go beyond the level of tendencies for the organism's regulatory pathways. I believe that the problem is most researchers do so in "global" manners, instead of focusing on biological meanings behind the microarray data. In my research, I will apply computational methods to the gene expression data using biological knowledge to interpret the gene expressions.

＜研究計画・方法＞

In order to uncover the mysteries of regulatory networks, it is the utmost importance to classify the functionally-unknown genes into their unique groups of genes. Currently, many researchers use two approaches (pattern discovery and class prediction) and their results are evaluated by the functional categories, the upstream sequence similarities, and the motif sequence similarities. The difference between the current approaches and my approach is the size and the number of datasets. Instead of using the microarray data as a large dataset, I will divide it into smaller datasets using biological knowledge. It is common practice to use the functional categories provided by Munich Information Center for Protein Sequences (MIPS) and Gene Ontology to evaluate how good these functionally-unknown genes are classified into their unique groups. If these categories are used as an evaluating standard, I will use them as a starting point to divide the dataset.

After creating smaller datasets, I will use the k-means clustering to classify genes into their unique groups. Current problems with using clustering method are setting the right number for clusters and the consistency of the results. For the first problem, I will use k=30 and k=60 because these numbers are currently used often in microarray research. For the second problem, I will run the clustering program ten times for each dataset and take an average for each specific object. By doing this, I will eliminate any bias resulting from clustering method.

Next step will be to find rules among clustered groups. Currently, 141 regulatory genes are listed in the Yeast Proteome Database. Out of them, 124 have been shown through genetic and biochemical experiments to encode regulating proteins that can influence the expression of specific genes. Guelzim et al. (2002) studied these 124 genes extensively and made their list of regulatory pathways available to the public. To evaluate my clustered results, I will use this list as a standard. However, this list is not perfect and very limited amount of information is shown; therefore, I believe that any information that I might extract from these clustered results should add more knowledge to the regulatory pathways of the yeast. Also, I would like to further test any interesting rules that I find out from these clustered results by analyzing the time-course microarray data for the yeast, which are available from the Saccharomyces Genome Database, and comparing the upstream sequence similarities and the motif sequences among coexpressed genes.

＜現在までに単位取得した専門教科＞	概論：3科目、方法論：1科目、専門：1科目
K212 知識ベース方法論 B	K217 物理科学概論 B
K215 イノベーション概論 B	K417 知識創発論
K216 知識科学概論 A	K616 生命知識特論