Probabilistic Models for Clustering Cell Cycle-Regulated Genes in the Yeast

Hyung-Joo Shin, Jeong-Ho Chang, Jin-San Yang, Byoung-Tak Zhang, Sirk June Augh*

School of Computer Science and Engineering, Korea Research Institute of
Seoul National University, Korea. Bioscience and Biotechnology*
{hjshin, jhchang, jsyang, btzhang}@scai.snu.ac.kr, june@bionet.snu.ac.kr

We used a probabilistic clustering method to identify yeast genes whose transcript levels are cell cycle regulated. There are 6178 genes in yeast Saccharomyces cerevisiae and 104 of them are discovered to be cell cycle regulated. It was estimated that some 250 cell cycle regulated genes might exist. The problem is to identify the cell cycle-regulated genes and to cluster them into 5 clusters: M/G1 boundary, late G1, S-phase, S/G2 phase, G2/M phase.

To this end, we made a clustering model fitted to 104 known genes. First, we divide 104 known genes into 60% as training set and 40% as test set. And then, from the three independent experiments, -- alpha-factor arrest, elutriation, and arrest of a cdc15 temperature-sensitive mutant, as we have 3 time series data for each gene (which belongs to training set) -- we found the prototypes of each cluster for three experiments: 5 prototypes for each experiment. And then, given a test data gene, among three time series data from three experiments, we select one time series whose entropy is the highest (which means that it varies most as cell cycle proceeds). For the 5 prototypes of that experiments, we calculate similarities with that gene and assign it to the cluster whose prototype is the most similar to that gene. Here, we find an appropriate threshold of similarity to assign a gene to a cluster.

For the other genes, which are not known whether they are cell cycle regulated, we do the same procedure described above to identify cell cycle-regulated genes.