ABSTRACT

This thesis aims to investigate for suitable statistical techniques to be integrated as a systematic statistical process for the analysis of gene expression data under the motivation of encountering the potential problems of limited replication and unknown distribution typically found in the public data sets such as the case study, where the conventional techniques are not applicable. Three investigated techniques include data normalization, identification of significant genes, and conditional cluster analysis. In this case, intensity-dependent normalization employing *lowess* function is applied to reduce the effect of dye bias where nonparametric correlation analysis is suggested so as to overcome non-normal distribution of random variables.

To identify some genes presenting significant expression patterns across the experiments, analysis of variance is applicable on the replicates of individual genes accompanied by the cutoff *p*-value based on Bonferroni correction to collect highly significant genes. Then, the process utilizes the identified significant genes to infer expression patterns of the other genes clustered in the same groups by a suitable technique, which algorithm on *k*-medoids is recommended as it does not rely on any assumption and also robust to the outliers. Extensively, principal component analysis is possibly used so as to explore expression patterns of some more genes nearby the output clusters.

The proposed systematic process was applied to the case study, which consists of two data sets yielded from cDNA microarrays in monitoring of biological response to the temperature changes of the baking yeasts, *Saccharomyces cerevisiae*. The dye bias

was reduced using intensity-dependent normalization that allows comparison of spots across the slides as demonstrated by the *MA* plots. The normalized data was analyzed to identify highly significant genes consequently utilized to infer expression patterns of the other genes in the same groups resulted by conditional clustering based on three steps of *k*-medoids clustering using CLARA and PAM. The output clusters consist of genes with expression patterns corresponding to their biological functions typically concerned to housekeeping genes. Apart from this, the use of principal components can explore expression patterns of some genes nearby the highlighted output clusters.