Gene Expression Data Analysis Using Machine Learning Techniques

Abdullah Mohammed Abdullah Khamis

Supervised By
Prof. Radi Abdel-Rahman Teleb
Dr. Mohammed H. Al-Qahtani

Abstract

The advent of highly parallel genomic platforms like microarray technology has facilitated a principal transition from gene science to genome science. Microarray provides the researchers with enormous data for the expression levels of thousands of genes measured at once. Full exploitation of microarray data aims to explore the complex relationships between genes and other regulatory network components that underpin all biological processes. However, microarray data contains little information about how these genes are regulated, and needs sophisticated machine learning methods to achieve this task.

Quantitative estimation of the regulatory relationship between transcription factors and genes is a key problem when trying to model the gene regulatory network. Because the difficulty of measuring the concentration levels of transcription factors and the fact that transcription factors are post-transcriptionally regulated, most of the work in the literature has been dedicated to infer the regulatory relationship between the transcription factors and the target genes from the expression levels of these genes.

In this work, a novel probabilistic model based on Gaussian Mixture Regression (GMR) is presented to derive the relationship between the transcription factor protein and each of its target genes quantitatively by estimating the sensitivity of these genes to the transcription factor's activity; and to generate a ranked list of predicted genes.

The potential power of GMR lies in its structure that combines the advantages of parametric models and the flexibility of non-parametric models. The GMR model is trained using microarray time series data and the verification scores, which are inferred using small interfering RNA. The model is applied to the p53 network and the results are compared with similar work based on differential equations. The investigated model provides us with a practical toolkit that can be applied to other transcription factors. The obtained information can be further exploited to build intelligent cancer classification systems.