Project 2: Microarray Data Analysis

Introduction

In the past few years, microarray technology has become one of the foremost tools in biological research. The emergence of this technology has empowered researchers in functional genomics to monitor gene expression profiles of thousands of genes (perhaps even an entire genome) at a time. However, mining microarray data also presents great challenges to Bioinformatics research. This project will acquaint you with several basic approaches to analyzing microarray data from the beginning to end. You will apply the techniques introduced in class to real-world microarray data sets and learn how to discover useful knowledge from the data sets. This project will also help you understand the challenges in microarray data analysis and motivate you to develop novel approaches to addressing those challenges.

Project Description

Part I: Time Series Data Analysis

In this part, you are required to choose one clustering algorithm from each category of the approaches introduced in class (partition-based, hierarchical, and density-based) to find clusters of genes which exhibit similar expression profiles. Give the pros and cons of existing methods. If you can design your own new algorithm which performs better than other methods for time series microarray data sets, you don't have to compare the existing algorithms, but you need to compare your algorithm with at least one of the existing algorithms. Your group also needs to give a presentation on your new algorithm on Nov.30 in class.

Part II:

Class prediction: design/implement an algorithm for classifying samples. This includes selecting informative genes as features and dividing samples into classes. If you can design your own new algorithm, your group needs to give a presentation on your new algorithm on Nov.30 in class.

Part III: Visualization

Visualization of microarray data provides users a direct impression of the data distribution. By visualization, users may gain some intuition regarding the relationships among data objects and the intrinsic structure of the data set. Visualization is especially important in the early stages of data analysis in which qualitative analysis is primary to quantitative. In this part, you are required to implement a high dimension to two dimension mapping to visualize the given microarray data sets. This may be used to check the results from your clustering/classification. Your group may give a presentation on your new algorithm on Nov.30 in class.

Part IV: Validation

In this part, you are required to validate your clustering results with the following methods:

- Choose an external index and compare the clustering results from different clustering algorithms with an external index (the ground truth clusters will be provided).
- Choose an internal index and compare the clustering results.

- Use the internal index you choose in the above step, and apply the statistical framework introduced in class to test the reliability of your clustering results.
- Using visualization to validate your results: dye the data objects in different clusters with different colors in your visualization. How does the data set look like? Does the colored visualization suggest good clustering?

Implementing basic existing algorithms will guarantee a B+ grade. Implementing more advanced algorithms or designing new algorithms will be considered for A grade.