Chapter 3

Building Gene Networks by Analyzing Gene Expression Profiles

Crescenzio Gallo
University of Foggia, Italy

ABSTRACT

The possible applications of modeling and simulation in the field of bioinformatics are very extensive, ranging from understanding basic metabolic paths to exploring genetic variability. Experimental results carried out with DNA microarrays allow researchers to measure expression levels for thousands of genes simultaneously, across different conditions and over time. A key step in the analysis of gene expression data is the detection of groups of genes that manifest similar expression patterns. In this chapter, the authors examine various methods for analyzing gene expression data, addressing the important topics of (1) selecting the most differentially expressed genes, (2) grouping them by means of their relationships, and (3) classifying samples based on gene expressions.

INTRODUCTION

Since the detection of the composition of DNA our understanding of biological structures and processes has expanded to a great extent, mostly thanks to computer science which plays a fundamental role in the field of bioinformatics. The main target at present is to analyze and employ the huge amount of accessible data. It is particularly important to distinguish various diseases through useful selection of gene indicators for morbid state and information about the possible correlations between genes.

Data analysis is seen as the largest and possibly the most important area of microarray bioinformatics to obtain the above said targets. Some specific data analysis methods address the fundamental scientific questions about microarray data, that is:

1. Which genes are differentially expressed in one set of samples relative to another,
2. What are the associations between the genes or samples being observed, and
3. Is it possible to group samples based on gene expression values?

DOI: 10.4018/978-1-5225-7489-7.ch003

In the next section, we illustrate the basic concepts underlying the previous questions and the bioinformatics research. Then we describe the methods for the first of these questions: the search for differentially (up or down) expressed genes. The following sections address the other two topics of clustering and classifying gene profiles. In the end, we show some concerns and issues of interest for future study and development in the field of (microarray) bioinformatics.

BACKGROUND

Gene expression profiling is an extensively used method in the analysis of microarray data. The leading hypothesis is that genes with similar expression profiles are co-regulated and are probably connected functionally.

Cluster analysis helps reaching these objectives; in particular, gene expression clusters help typify unknown genes assigned to the cluster by those genes that have a known function, and are the support for distinguishing common upstream regulatory sequence elements (Brazma *et al*., 2000).

Clustering of expression profiles and functional grouping is especially compelling if the complete gene set is known. Hence, we used the large publicly available data set included in the Stanford Yeast Database at http://genome-www.stanford.edu for our clustering study.

Many applications aim at the molecular classification of diseases based on gene expression profiling and clustering. See, for example, works on leukemias (Golub *et al*., 1999) and B-cell lymphomas (Alizadeh *et al*., 2000). These and other studies confirm the usefulness of microarray bioinformatics for scientific and industrial research.

Gene Expression Profiling

Having M array probes with N samples (time points, patient tissues, *etc*.), you construct an $M \times N$ data matrix (the *gene expression matrix*), where the M rows describe the gene expression values across the experiments and the N columns (samples) describe the experiments across the gene set. Practically, each gene is assigned a set of (possibly normalized) numerical values (the *gene expression profile*) corresponding to the gene's "presence" in each sample.

Then, an $M \times M$ similarity matrix is obtained by calculating the "closeness" between each gene pair with values inversely proportional to the relative (expression profile) distance. Typically Pearson correlation, Spearman's rank correlation, Hamming distance, Euclidean distance and mutual information are employed as similarity measures (Jain & Dubes, 1988; Mirkin, 1996), each having its specific advantages and disadvantages.

The (normalized and possibly log-transformed) expression profiles of thousands of genes are first examined under the typical two-fold comparison (between the same patients in *control vs treatment*, or between two different groups of patients), in order to identify differentially expressed genes. This can be done using an appropriate statistic (t-statistic, Wilcoxon, Mann-Whitney depending on the type of comparison) to compare gene expression variability.

Then selected genes are clustered in order to find groups of co-regulated genes.

The clustering output in the end serves as the basis for typifying the role of undetermined genes by means of information available from known genes, to recognize supposed regulatory elements in the

Related Content

T-Scan Case Finishing Applications in Aesthetic Dentistry Assisted by the iTero Digital Impression System: A Case Report

Christopher J. Stevens, DDS (2015). *Handbook of Research on Computerized Occlusal Analysis Technology Applications in Dental Medicine (pp. 705-743).*

www.irma-international.org/chapter/t-scan-case-finishing-applications-in-aesthetic-dentistry-assisted-by-the-itero-digital-impression-system/122086

Adding Technology to Diagnostic Methods

John C. Radke, BM, MBA (2015). *Handbook of Research on Computerized Occlusal Analysis Technology Applications in Dental Medicine (pp. 153-214).*

www.irma-international.org/chapter/adding-technology-to-diagnostic-methods/122073

Multi-Criteria Decision-Making Techniques for Histopathological Image Classification

Revathi T., Saroja S., Haseena S.and Blessa Binolin Pepsi M. (2019). *Histopathological Image Analysis in Medical Decision Making (pp. 103-138).*

www.irma-international.org/chapter/multi-criteria-decision-making-techniques-for-histopathological-image-classification/212541

How to Educate Health Organizations to Apply Lean Methodology

Dulce Lourenço Mirandaand Carlos Alberto Silva (2022). *Handbook of Research on Improving Allied Health Professions Education: Advancing Clinical Training and Interdisciplinary Translational Research (pp. 333-344).*

www.irma-international.org/chapter/how-to-educate-health-organizations-to-apply-lean-methodology/302533

Emerging Role of Artificial Intelligence and Robotics in Physiotherapy: Past, Present, and Future Perspective

V. Rajalaxmiand R. Kirthika (2024). *Advancements in Clinical Medicine (pp. 200-214).*

www.irma-international.org/chapter/emerging-role-of-artificial-intelligence-and-robotics-in-physiotherapy/346201