Building Gene Networks by Analyzing Gene Expression Profiles

Crescenzio Gallo
University of Foggia, Italy

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?

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

DOI: 10.4018/978-1-5225-2255-3.ch039

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 early regions of the genes in the same clusters and reduce redundancy and generate averaging profiles in the context of regulatory networks.

Cluster Analysis

Cluster analysis (Duda *et al.*, 2001; Jain & Dubes, 1998; Jain *et al.*, 1999) can be summarized as follows. Let n experimental outcomes $x_i \in \mathbb{R}^m$ ($i=1..n$; each point has m components): the objective is to identify the underlying structure of the data, partitioning the n points into k clusters in order to group in the same cluster points "closer" to each other than to points belonging to different clusters.

In the above statement no clear definition exists for "closer" points, and this depends on the resolution at which the data are viewed. The last issue is typically addressed by generating a tree of clusters (a dendrogram), whose number and structure depend on the resolution that is used.

Two of the most common methods of clustering gene expression data are hierarchical clustering and $k-$means clustering (Geraci *et al.*, 2009; Jain *et al.*, 1999).

Hierarchical clustering is the most used, and produces a representation of the data with the most similar patterns grouped in a hierarchy of subsets. This method, however, suffers from considerable problems when applied to data containing a significant amount of *noise*, revealing itself of low applicability. In this case the solutions may not be unique and be data-order dependent. Mathematically, hierarchical clustering involves computing a matrix of all distances for each expression measurement in the study, merging and averaging the values of the closest nodes, and repeating the process until all nodes are merged into a single node.

$K-$means clustering involves generating cluster centers in $n-$dimensions and computing the distance of each data point from each of the cluster centers. The data points are assigned to their closest cluster center. A new cluster position is then computed by averaging the data points assigned to the cluster center. The process is repeated until the positions of the cluster centers stabilize.

UNSUPERVISED NEURAL NETWORKS

Artificial Neural Networks can be used not only for prediction but also for data classification. Unlike regression problems, where the goal is to produce a particular output value for a given input, classification problems require labeling of all data as belonging to one of n known classes. These classification models are typical cases of supervised networks, in which it is a priori possible to associate data to clusters and to train the neural network for the classification of further data.