Computers & Operations Research 39 (2012) 3046–3061

Contents lists available at SciVerse ScienceDirect

Computers & Operations Research

journal homepage: www.elsevier.com/locate/caor

Review

Clustering of high throughput gene expression data

Harun Pirim [a,*], Burak Ekşioğlu [a], Andy D. Perkins [b], Çetin Yüceer [c]

[a] Department of Industrial and Systems Engineering, Mississippi State University, P.O. Box 9542, Mississippi State, MS 39762, United States
[b] Department of Computer Science and Engineering, Mississippi State University, United States
[c] Department of Forestry, Mississippi State University, United States

ARTICLE INFO

Available online 29 March 2012

Keywords:
Clustering
Bioinformatics
Gene expression data
High throughput data
Microarrays

ABSTRACT

High throughput biological data need to be processed, analyzed, and interpreted to address problems in life sciences. Bioinformatics, computational biology, and systems biology deal with biological problems using computational methods. Clustering is one of the methods used to gain insight into biological processes, particularly at the genomics level. Clearly, clustering can be used in many areas of biological data analysis. However, this paper presents a review of the current clustering algorithms designed especially for analyzing *gene expression* data. It is also intended to introduce one of the main problems in bioinformatics – clustering gene expression data – to the operations research community.

Contents

1. Introduction

Clustering in biology has a history that goes back to Aristotle's attempt to classify living organisms [6]. Today, clustering genomic data stands out as an approach to deal with high-dimensional data produced by high throughput technologies such as *gene* expression *microarrays* [84]. Biological data were limited to DNA sequence data before the *genome* age in the 1980s [68]. Nowadays, terabytes of high throughput biological information are generated with the advent of new technologies, such as microarrays, *eQTL* mapping, and *next generation sequencing*. Now, a need for exploiting computational methods exists to analyze and process such amounts of data in depth and in different ways to address complex biological questions regarding gene functions, gene co-expression, protein–protein interactions (PPI), personalized drug design, systems level functional analysis of plants and

* Corresponding author. Tel.: +1 662 325 4226.
E-mail address: harunpirim@gmail.com (H. Pirim).

0305-0548/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.cor.2012.03.008

animals, and organism–environment interaction. This fact has given birth to disciplines like bioinformatics, computational biology, and *systems biology*.

In physics, before mathematical models were incorporated, i.e., before Newton, the discipline was stamp collecting (i.e., descriptive). Incorporation of mathematical models changed physics into a predictive science. In a similar manner, incorporation of computation into biology is changing the discipline from being a descriptive science to a predictive science. One of the prediction methods used in biology to analyze the high throughput data is clustering. As a data mining method, clustering of gene expression data was well studied during the last decade. Clustering is also a well-known and studied problem in the operations research (OR) field. However, clustering of gene expression data is not extensively studied by the OR community, although data mining techniques have been used in market segmentation and facility location problems.

Certain aspects of biological theories can be modeled using OR tools. One of these aspects is that a small subset of genes are typically involved in a particular cellular process of interest, and a cellular process happens only in a subset of the *samples* [65]. Another aspect is that genes of the same pathway may be induced or suppressed simultaneously or sequentially upon receiving stimuli [149]. A third aspect is that most biologists assume an approximately *scale-free topology*, or a *small world property*, for graphs constructed from gene expression data [145]. Hence, one may say that genes with high *connectivity* are much fewer in number than genes with low connectivity [131]. Thus, this review discusses many diverse approaches and algorithms that currently exist for clustering of gene expression data from an OR perspective by introducing background in molecular biology, and presenting clustering approaches and techniques. The paper is organized as follows: Section 2 gives concise information about molecular biology and relevant disciplines; Section 3 provides a problem definition for clustering gene expression data as well as representations of expression data; Section 4 reviews recent algorithms used for clustering gene expression data; Section 5 suggests future research directions for the operations research community; and Appendix A provides the glossary that includes definitions of the italicized words and phrases throughout the text.

2. Biological background

The essential cellular molecules for a biological system to function and interact with its surrounding include DNA, RNA, proteins, and *metabolite*s, all of which are under physiological and environmental control. Many different interaction layers exist among these molecules such as PPI networks, i.e., interactomes, gene regulatory networks (GRNs), biochemical networks, and gene co-expression networks. A holistic picture of these interactions is being studied through systems biology.

Based on the central dogma of molecular biology, DNA transcribes into RNA, and RNA translates into proteins, some of which then serve as catalysts in the production of metabolites. A gene is expressed upon receiving the transcriptional signal. Genes have *activator*s and *repressor*s. Genes reveal no or low expression values without activators. Repressors block gene expression, even in the presence of activators. *Transcription factor*s (TFs) are activator or repressor proteins produced by genes. TFs bind to *regulatory site*s and turn them on to transcribe RNA or off. Genes may show cascade interactions. For example, the product of one gene may increase or decrease the transcription rate of the other, and this process may continue downstream including temporal or causal order of molecular events.

Fig. 1. A microarray chip produced by Affymetrix courtesy. (*source*: http://www.affymetrix.com/about_affymetrix/media/image-library.affx).

It is often preferred to analyze thousands of genes' dynamics together rather than one at a time. The DNA microarray (Fig. 1) has been one of the commonly used technologies to measure thousands of gene expressions simultaneously [84], and microarray data have been stored in public databases such as the Gene Expression Omnibus (GEO) for further analysis. For example, the Affymetrix GeneChip Mouse Genome 430 2.0 Array provide 45,000 probe sets to analyze expression levels of more than 39,000 transcripts. Its *Feature* size is 11 μM. Eleven probe pairs per sequence are used.

The data extracted from microarrays or a similar technology is analyzed using a *reverse engineering* approach. A simplified framework of reverse engineering methodology for modeling GRNs from gene expression data is shown in Fig. 2, which is adapted from Lee and Tzou [76]. However, it is a challenging task to infer about GRNs because expression data are high-dimensional, complex, and non-linear. Further complicating the inference are the facts that, dynamic relations exist among thousands of genes, expression data involve *noise*, and the sample-to-gene ratio is normally very small [147] because the array chips corresponding to samples are expensive. Co-expressed genes show coherent *expression pattern*s, indicating that they may have similar functions [84] or co-exist in a pathway. However, different external conditions may trigger a gene to be expressed similarly with different group of genes [84]. Genes with similar expression patterns are more likely to regulate each other or to be regulated by a parent gene [92]. Here, the problem of quantifying the relations between genes arises.

A powerful clustering approach as well as a predictive model may detect patterns or relationships in expression data [84]. However, a predictive model should be guided by biological facts, meaning that results of predictive models should be validated by biological knowledge. On the other hand, biological experiments should be guided by computational methods to make the best use of data and reduce experimental and time costs (Fig. 3). Online databases exist to facilitate *validation* of the results obtained from predictive models. Incorporation of the database knowledge to modeling GRNs is essential for more accurate results or for comparing the model to reality.

3. Problem definition and representations of gene expression data

Clustering generates individual groups of data called a *partition*, rather than assigning *objects* into already known groups as in *classification* [8]. A partition is defined as follows:

$P=\{c_1,c_2,\ldots,c_s\}$, where s is the number of clusters; $\sum_{i=1}^{s}|c_i|=n$, where n is the number of objects; and $|c_i|$ is the cardinality of *cluster i*.

$X=\{x_1,x_2,\ldots,x_n\}$ is the set of n objects and $Y=\{y_1,y_2,\ldots,y_n\}$ is the set of n patterns, where $y_i \in R^d$ and d is the number of samples.