Neurocomputing 73 (2010) 1425–1437

Contents lists available at ScienceDirect

Neurocomputing

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

Efficiently mining local conserved clusters from gene expression data

Guoren Wang *, Yuhai Zhao, Xiangguo Zhao, Botao Wang, Baiyou Qiao

College of Computer Science and Engineering, Northeastern University, Shenyang 110004, China

ARTICLE INFO

Article history:
Received 17 April 2009
Received in revised form
15 September 2009
Accepted 29 November 2009
Communicated by L. Kurgan
Available online 22 December 2009

Keywords:
Bioinformatics
Clustering
Gene expression data

ABSTRACT

Extensive studies have shown that mining gene expression data is important for both bioinformatics research and biomedical applications. However, most existing studies focus only on either co-regulated gene clusters or emerging patterns. Factually, another analysis scheme, i.e. *simultaneously mining phenotypes and diagnostic genes*, is also biologically significant, which has received relative little attention so far. In this paper, we explore a novel concept of local conserved gene cluster (LC-Cluster) to address this problem. Specifically, an LC-Cluster contains a subset of genes and a subset of conditions such that the genes show steady expression values (*instead of the coherent pattern rising and falling synchronously defined by some previous work*) only on the subset of conditions but not along all given conditions. To avoid the exponential growth in subspace search, we further present two efficient algorithms, namely FALCONER and E-FALCONER, to mine the complete set of maximal LC-Clusters from gene expression data sets based on enumeration tree. Extensive experiments conducted on both real gene expression data sets and synthetic data sets show: (1) our approaches are efficient and effective, (2) our approaches outperform the existing enumeration tree based algorithms, and (3) our approaches can discover an amount of LC-Clusters, which are potentially of high biological significance.

1. Introduction

Clustering gene expression data is an important research topic in bioinformatics [1–4]. Compared to the traditional clustering approaches, biclustering algorithms, which perform simultaneous row–column clustering in a data matrix, are more suitable for discovering local expression patterns, formed by a subset of genes across a subset of samples, from gene expression data.

A series of biclustering methods have been proposed in the past decade. They all aim to capture "blocks", also called biclusters, within gene expression matrices, however, since different model definitions, the internal characteristics of "blocks" vary from model to model. "Blocks" of different structures imply different biological significance, and thus the corresponding biclustering algorithms should serve different analyzing purposes of gene expression data. As such, in this paper, we propose a novel model, namely LC-Cluster, serving the purpose of *simultaneous mining phenotypes and diagnostic genes*, as also known as the cell phenotype prediction problem [5,6].

Unlike the coherent pattern rising and falling synchronously defined by some previous work [7–10], given a gene g_i, we say it to be a *diagnostic gene* if it shows *steady expression values* only across a proper subset of samples, say S, instead of all or a large majority of the samples. We also say g_i is *local conserved* across S, which may correspond to a specific phenotype. For example, Fig. 1 gives a simplistic illustration of the gene expression patterns via a data set of three phenotypes, labeled as "Phenotype 1", "Phenotype 2", and "Phenotype 3", respectively. *Different from the previous work*, which focus on co-regulated gene clusters or emerging patterns, our primary goal of analyzing such data sets is to discover the three classes of the samples while identifying some subsets of genes manifesting this class structure without any priori knowledge. Specifically, in Fig. 1, $gene_1$ is a *perfect diagnostic gene* and $gene_2$ is an *approximate diagnostic gene* since the expression levels of $gene_1$ are equally low for Phenotype 1, equally high for Phenotype 2 and equally intermediate for Phenotype 3, however, all cases for $gene_2$ are approximately.

Obviously, it provides quite valuable hypothesis for biologists to identify such a group of genes and samples since the samples in the same cluster probably indicate a specific phenotype while the genes may suggest all candidates related to the phenotype [5,6,11]. We call such a cluster a *local conserved gene cluster* or an *LC-Cluster*. Note: at first sight, our investigation is similar to some early work, such as some existing bicluster models [7–10,12] or emerging pattern [13,14], but there exist significant inherent differences between them indeed. The detailed

* Corresponding author.
E-mail address: wanggr@mail.neu.edu.cn (G. Wang).

0925-2312/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2009.11.009

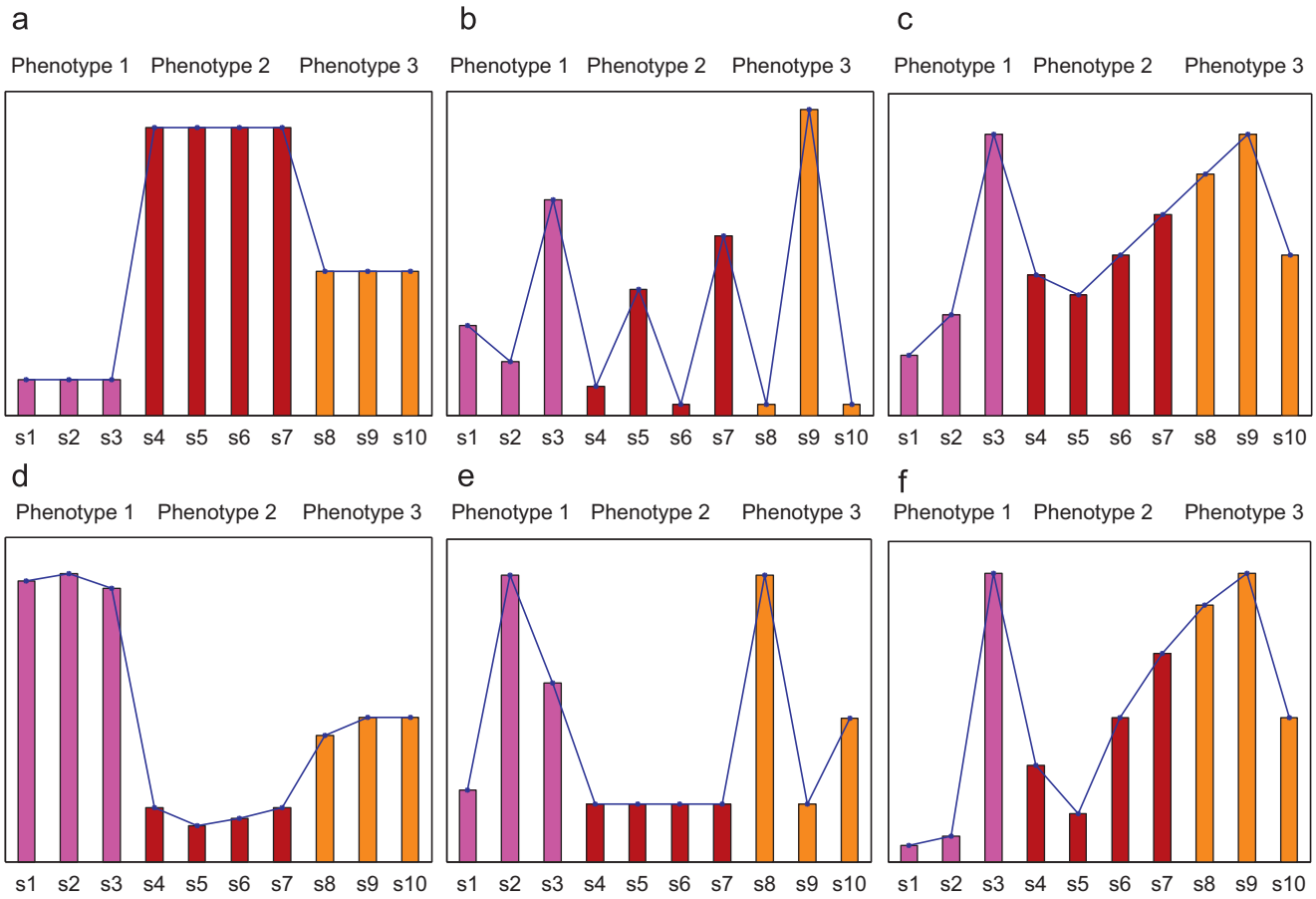

Fig. 1. Examples of the gene expression patterns across three phenotypes. The first three samples, i.e. s_1-s_3, belong to Phenotype 1, the second four samples, i.e. s_4-s_7, to Phenotype 2 and remainder, s_8-s_{10}, to Phenotype 3. (a) $gene_1$. (b) $gene_3$. (c) $gene_5$. (d) $gene_2$. (e) $gene_4$. (f) $gene_6$.

explanations crucial to make sense of these differences and comprehend our work are as follows.

Difference between existing biclusters and LC-Cluster: The concept of bicluster was introduced by Hartigan [15] and first applied to gene expression data analysis by Cheng and Church [12]. Proven very useful at uncovering hidden local structures in gene expression data [16,17], many biclustering approaches have been proposed. Some are based on heuristic approaches: Cheng and Church [12], OPSM [18,16], $\delta-$cluster [19] and FLOC [20], etc. Some are based on tree-structure: p-cluster [7,21], OP-Cluster [22], Maple [23] and SeqClus [21], etc. Some are based on graph theory: CAST [24], CLICK [25] and SAMBA [26], etc. There are also some bicluster-based variants: Plaid model [27], Spectral model [28], etc.

As aforementioned, different bicluster models should serve different analyzing purposes of gene expression data, since different kinds of resulting clusters are of the same "block" appearance but of *different internal characteristics*. To differentiate the proposed LC-Cluster from other numerous biclusters, an elaborate comparison of two representative bicluster models, i.e. the model of Cheng and Church [12] (called CC-Cluster) and p-Cluster [7], and LC-Cluster is given below.

The pioneering CC-Cluster [2] introduced a measure of *mean squared residue*, H, to assess the overall homogeneity of a bicluster (I,J), which, provably (cf. ANOVA), can be further uniquely defined by the bicluster variance, $(1/|I||J|)\sum_{i\in I}\sum_{j\in J}(d[i,j]\text{-}d[I,J])^2$, minus the variance of row mean, $(1/|I|)\sum_{i\in I}(d[i,J]\text{-}d[I,J])^2$, minus the variance of column mean, $(1/|J|)\sum_{j\in J}(d[I,j]-d[I,J])^2$:

$$H(I,J)=\frac{1}{|I||J|}\sum_{i\in I,j\in J}(d[i,j]-d[i,J]-d[I,j]+d[I,J])^2=\frac{1}{|I||J|}\sum_{i\in I}\sum_{j\in J}(d[i,j]\text{-}d[I,J])^2-\frac{1}{|I|}\sum_{i\in I}(d[i,J]\text{-}d[I,J])^2-\frac{1}{|J|}\sum_{j\in J}(d[I,j]-d[I,J])^2 \quad (1)$$

The preceding formula provides deep insight into the nature of CC-Cluster model, which indicates that a bicluster of low mean squared residue corresponds to a given submatrix of well-separated row means and/or well-separated column means. This makes intuitive sense, since the well-separated row means indicate data is well-organized along rows, the well-separated column means indicate data is well-organized along columns, and both well-separated row means and well-separated column means indicate data is well-organized within the whole bicluster. However, *what Eq. (1) measures is only a macroscopic coherence, and thus a submatrix of large row variance or large column variance could also be a possible bicluster captured by the approach of Cheng and Church [12].* Note: large row variance means the elements in the corresponding subset of samples are unlikely to have the similar properties, although it does not affect a perfect bicluster generated. Consider a special case illustrated in Fig. 2(a), three genes, i.e. g_1, g_2 and g_3, each of large row variance but the same pure shifting pattern, form a perfect bicluster of the lowest mean squared residue, 0, on the given subset of samples, where the samples could be different phenotypes. Moreover, patterns