# Biclustering Gene-Feature Matrices for Statistically Significant Patterns

Mehmet Koyutürk, Wojciech Szpankowski, and Ananth Grama
Dept. of Computer Sciences, Purdue University
West Lafayette, IN 47907
{koyuturk,spa,ayg}@cs.purdue.edu

## Abstract

*Biclustering is an important problem that arises in diverse applications, including analysis of gene expression and drug interaction data. The problem can be formalized in various ways through different interpretation of data and associated optimization functions. We focus on the problem of finding unusually dense patterns in binary (0-1) matrices. This formulation is appropriate for analyzing experimental datasets that come from not only binary quantization of gene expression data, but also more comprehensive datasets such as gene-feature matrices that include functions of coded proteins and motifs in the coding sequence.*

*We formalize the notion of an "unusually" dense submatrix to evaluate the interestingness of a pattern in terms of statistical significance based on the assumption of a uniform memoryless source. We then derive simplified formulations to assess statistical significance of discovered patterns. Using statistical significance as an objective function, we formulate the problem as one of finding maximally significant dense submatrices of a large sparse matrix. Adapting a simple iterative heuristic along with randomized initialization techniques, we derive fast algorithms for discovering binary biclusters. We conduct experiments on a binary gene-feature matrix and a quantized breast tumor gene expression matrix. Our experimental results show that the proposed method quickly discovers all interesting patterns on these datasets.*

## 1. Introduction

Biclustering, which has been applied intensively in molecular biology research recently, provides a framework for finding hidden substructures in large high-dimensional matrices. In general, the problem can be defined as one of finding large sets of rows and columns such that the rows show unusual similarities along the dimensions characterized by columns and vice versa. It is possible to define these similarities in terms of correlation of gene expression vectors [2], preserving the order of expression levels [1], matching symbols over a finite alphabet [4] or high density of gene expression or features [6]. In order to understand and interpret the biological relevance of discovered patterns, any of these formulations need to be associated with a formulation of statistical significance of biclusters.

In this paper, we address the problem of finding unusually dense submatrices in a binary matrix. Binary matrices can arise from quantization of gene expression data [5] or more comprehensive datasets such as gene-feature matrices that include functions of coded proteins and motifs in the coding sequence. We formalize the notion of an "unusually dense" submatrix in the next section, formulate statistical significance as an objective function for an optimzation problem and develop fast heuristics to solve this problem in Section 3. We present experimental results on the performance of these formulation and algorithms in Section 4.

## 2. Statistical Significance of Biclusters

Given binary matrix $G$ with $M$ rows, $N$ columns and $K$ ones, we are looking for a subset of rows and columns such that the submatrix induced by these rows and columns is dense enough to be considered statistically significant. Here, the rows and columns of the submatrix do not have to be contiguous. We assume that the matrix is generated by a uniform memoryless source, with $Pr\{G(i,j) = 1\} = p$. $p$ can be estimated by the density of the matrix, *i.e.*, $p = K/MN$. Now, for an arbitrary set of $m$ rows and $n$ columns, assume that the number of ones in the corresponding submatrix is $k$. Then, since the number of ones in an arbitrary submatrix is from a binomial distribution with parameters $mn$ and $p$, using Chernoff's bound we get

$$Pr\{k \geq mnp(1+\epsilon)\} \leq e^{-mnp\epsilon^2/3} \quad (1)$$

for $\epsilon > 0$.

Now assume that we are interested in discovering all submatrices such that the probability of observing $k$ ones in the matrix is less than $P^*$. Then, by Equation 1, the observed

bicluster is significant if

$$e^{-mnp\epsilon^2/3} \leq P^* \,. \tag{2}$$

Then,

$$mnp\epsilon^2/3 \geq -\ln(P^*) \,. \tag{3}$$

Solving this equation and letting $E = -\ln(P^*)$, we find that a submatrix of $m$ rows, $n$ columns and $k$ ones is significant if $k \geq mnp(1+\epsilon)$, where

$$\epsilon \geq \sqrt{3E/mnp} \,. \tag{4}$$

In other words, in order to be considered as a significant bicluster, the number of ones in a submatrix has to deviate from the mean by at least $\sqrt{3Emnp}$. We use this result to derive an objective function for our randomized heuristic to find statistically significant biclusters.

## 3. Randomized Heuristics for Discovering Significant Biclusters

Following the results of the previous section, a submatrix of $m$ rows, $n$ columns and $k$ ones is a statistically significant bicluster if

$$C(m,n,k) = k - mnp - \sqrt{3Emnp} \geq 0 \tag{5}$$

Observing that a larger value of $C(m,n,k)$ means a more interesting bicluster, we consider $C(m,n,k)$ as an objective function and search for submatrices for which this function has a local maximum. For this purpose, we adapt an iterative heuristic based on alternating projections between row and column spaces.

Given a set of $m$ rows, consider the problem of choosing a set of columns to maximize $C(m,n,k)$. Define $x$ to be an $M$-dimensional binary vector where $x(i) = 1$ if and only if the $i^{th}$ row of $G$ is in the submatrix. Define $N$-dimensional binary vector $y$ similarly for the column space. Then, clearly the submatrix induced by $x$ and $y$ has $k = x^T G y$ ones. The following lemma provides a solution for $y$ that maximizes $C(|x|,|y|,x^TGy)$ for a fixed $x$.

**Lemma 1** *Given vector $x$ , let $s = G^T y$. Let $\Pi$ be a permutation of numbers from 1 to $N$ such that $s(\Pi_i) \geq s(\Pi_j)$ for $i < j$. Then, the vector $y$ that maximizes $C(|x|,|y|,x^TGy) = x^TGy - p|x||y| - \sqrt{3Ep|x||y|}$ is given by*

$$y(\Pi_i) = \begin{cases} 1 & if\ s(\Pi_i) \geq p|x| + \sqrt{3Ep|x|}(\sqrt{i} - \sqrt{i-1}) \\ 0 & otherwise \end{cases} \tag{6}$$

*for $1 \leq i \leq N$.*

The above lemma can be proven by simple algebra observing that if $y(i) = 1$ and $s(i) < s(j)$, then we must have $y(j) = 1$ for $C(|x|,|y|,x^TGy)$ to be maximal and the contribution of each column to this objective function is the difference between the number of ones of this column in the rows characterized by $x$ and the quantity on the right-hand-side of Equation 6. This lemma also applies to the solution of $x$ for a fixed $y$.

FINDBICLUSTER($G$, $E$)
▷ $G$ : Binary matrix
▷ $E$ : Desired level of significance ($-\ln(P^*)$)
1 initialize $y$ to a random binary vector
2 **repeat**
3 solve for $x$ to maximize $C(|x|,\ |y|,\ x^TGy)$
4 solve for $y$ to maximize $C(|x|,\ |y|,\ x^TGy)$
5 **until** no improvement on $C(|x|,\ |y|,\ x^TGy)$ is possible
6 **return** submatrix induced by $x$ and $y$ **if** it is significant

**Figure 1. Iterative algorithm for discovering significant biclusters in binary gene expression data for a desired level of significance(E).**

Observing that the algorithm provided by Lemma 1 only requires a sparse matrix-vector multiplication which can be performed in $O(K)$ time, we derive an iterative algorithm [3] for finding a maximally significant bicluster, which is shown in Figure 1.

Different runs of FINDBICLUSTER will converge into different local maxima since the initialization is random. This provides us with two possible methods of using this algorithm for finding all interesting biclusters.

1. Run FINDBICLUSTER several times to obtain a set of biclusters. Prune out the redundant ones among these, in terms of significance and overlaps between submatrices. Rank and return the remaining biclusters based on their significance.

2. Run FINDBICLUSTER several times to find a single bicluster that has maximum significance. Return this bicluster and filter out the submatrix associated with this bicluster. Repeat this procedure until no significant patterns can be found.

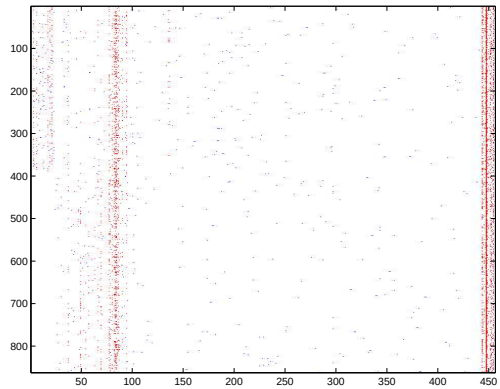
The first method has the advantage of discovering overlapping biclusters, which is very desirable but difficult to establish for many biclustering algorithms. However, as the second method filters out the most dominant pattern in the matrix, it makes possible for the algorithm to converge into less significant but still interesting biclusters.

## 4. Experimental Results

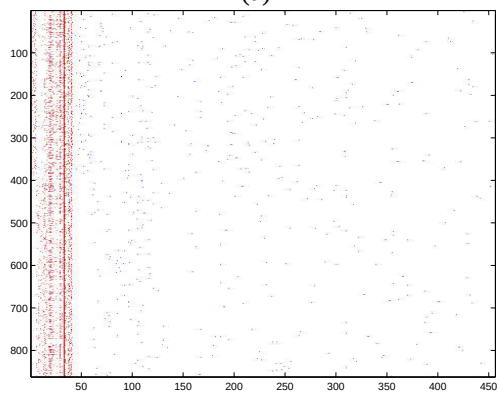
We first illustrate the performance of the proposed algorithm on a gene-feature dataset. This dataset was part of the KDD-Cup data mining competition in 2001[1]. It contains 862 genes from a particular organism. Each gene is associated with a set of features like protein classes, protein complexes, phenotypes, motifs and functions. Representing genes by rows and features by columns, we obtain a 862×456 binary matrix with 4339 ones (density:0.011), where a one signifies the association between the corresponding gene and feature (*e.g.*, if gene $i$ codes a protein that belongs to protein class $j$, then the $G(i,j) = 1$). The most dominant bicluster in this matrix is shown in Figure 2. In this figure, blue points show ones in the input matrix and red points show ones that are contained in the bicluster. This bicluster is composed of 42 functions/motifs and all genes, which has a density of 0.104, with a significance of $P^* \leq 10^{-6}$. This bicluster is a dominant local maximum of the objective function in Equation 5 and we observe that the algorithm almost always tends to converge to this solution. Thus, we conclude that this bicluster can be interpreted as a global pattern that contains common features in this particular organism. These features include protein classes like transcription factors, protein complexes like cytoskeleton, phenotypes like sensitivity to antibiotics, functions like cell growth, cell division and dna synthesis and a few motifs. In Figure 2(a), the first red block corresponds to protein classes, the second corresponds to phenotypes and the last one corresponds to functions. On the other hand, motifs are distributed sparsely across genes. Since this pattern is strongly dominant in this dataset, it is appropriate to filter out this submatrix as suggested by the second method of the previous section and rerun the biclustering algorithm to discover smaller interesting patterns. Doing so, we discover several small biclusters. As an example, one of these reveals that the protein class of tubulins is associated with two motifs, namely PS00227 and PS00227 with $P^* \leq 10^{-3}$.

We also conduct experiments on a gene expression dataset that is obtained from NCBI's GEO data collection [2]. This dataset contains gene expression data collected from 84 samples which are associated with several types of human breast cancer. The data is used for characterization of variation in gene expression in 65 surgical specimens of breast tumor from 42 individuals. The dataset contains 9216 probes.

We quantize the breast tumor data to signify whether each gene is significantly expressed or suppressed in each sample as follows.




(a)




(b)

**Figure 2. A strongly dominant bicluster in a gene-feature matrix. (a) Bicluster in the original binary matrix, (b) binary matrix reordered to bring rows and columns in bicluster together.**

$$E(i,j) = \begin{cases} 1 & if\ G(i,j) > \mu_j + \alpha\sigma_j \\ 0 & otherwise \end{cases} \quad (7)$$

$$S(i,j) = \begin{cases} 1 & if\ G(i,j) < \mu_j - \alpha\sigma_j \\ 0 & otherwise \end{cases} \quad (8)$$

Here, $G$ denotes the gene expression matrix, $\mu_j$ and $\sigma_j$ denote the mean and standard deviation of expression level across all genes for sample $j$, respectively. $\alpha$ is an input parameter that is used to tune the desired deviation from average. Binary matrix $E$ contains a one for each gene that is significantly expressed in the corresponding sample. Similarly, binary matrix $S$ signifies the supression of genes across samples. The experiments that are reported in this section are all conducted on the expression matrix ($E$), which is computed by setting $\alpha = 2$.

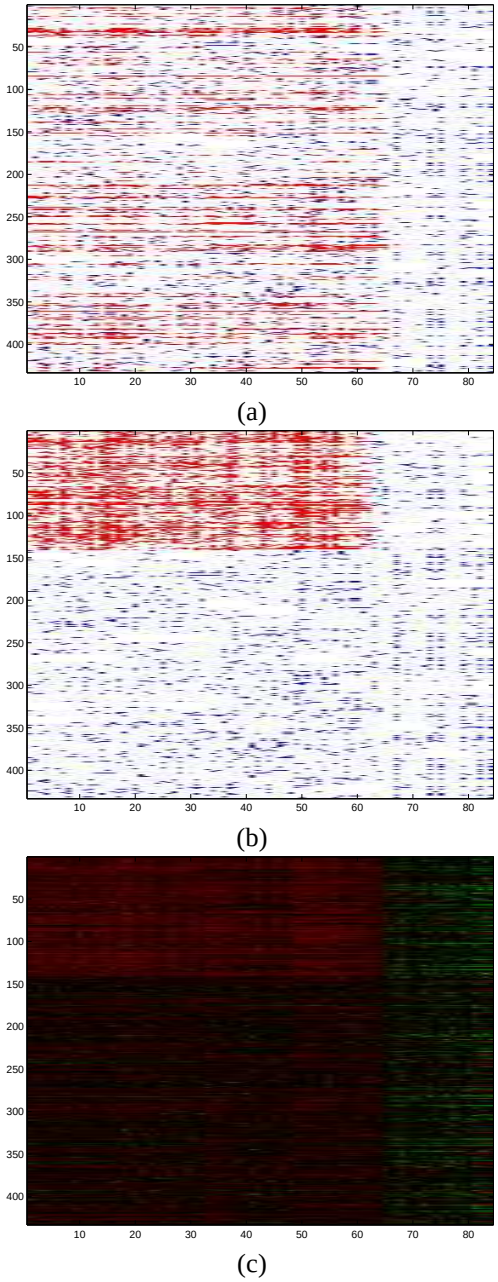
A dominant bicluster discovered in breast tumor data is shown in Figure 3. For the sake of visualization, since the

1 http://www.cs.wisc.edu/~dpage/kddcup2001/

2 http://www.ncbi.nlm.nih.gov/geo/gds/gds_browse.cgi

number of genes in the dataset is very large, we only show genes that are expressed in at least 3 samples as other genes are not interesting enough to construct significant biclusters. In Figure 3(a), the significant bicluster discovered on the binary expression matrix is shown. In Figure 3(b), the genes and samples that are contained in the bicluster are brought together to illustrate the density of the bicluster. The original gene expression matrix is reordered accordingly in Figure3(c) to illustrate that binary quantization and biclustering together are able to reveal a significant expression pattern in the breast tumor data. This bicluster contains 141 genes and 62 samples with a density of 0.37, while the entire matrix has a density of 0.077 ($P^* \leq 10^{-6}$). The samples in the bicluster all belong to luminal-like ER+ tumors, ERBB2+ tumors and basal-like tumors, while normal breast cells are left out.

Another significant bicluster that illustrates how local maxima of the objective function in Equation 5 is able to capture interesting patterns is shown in Figure 4. This bicluster contains 79 genes and 7 samples with a density of 0.56 ($P^* \leq 10^{-6}$) and is associated with normal breast cells. Note also that all biclusters presented in this section are discovered in less than a second by a simple implementation of the algorithm in Matlab on a 3GHz PC.




(a)

(b)

(c)

**Figure 3. A maximally significant bicluster discovered in breast tumor data. (a) Bicluster in the binary quantized matrix, (b) binary matrix reordered to bring rows and columns in bicluster together, (c) original gene expression matrix reordered accordingly.**

## 5. Conclusion

## References


[1] A. Ben-Dor, B. Chor, R. Karp, and Z. Yakhini. Discovering local structure in gene expression data: The order-preserving submatrix problem. In *Proc. 6th Annual Intl. Conf. Research in Computational Molecular Biology (RECOMB'02)*, 2002.

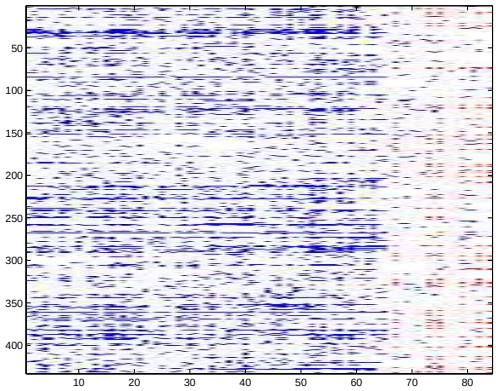
[2] Y. Cheng and G. M. Church. Biclustering of expression data. In *Proc. 8th Intl. Conf. Intelligent Systems for Molecular Biology (ISMB'00)*, pages 93–103, 2000.

[3] M. Koyutürk, A. Grama, and N. Ramakrishnan. Algebraic techniques for analysis of large discrete-valued datasets. In *Proc. 6th European Conf. Principles of Data Mining and Knowledge Discovery (PKDD 2002)*, pages 311–324, 2002.

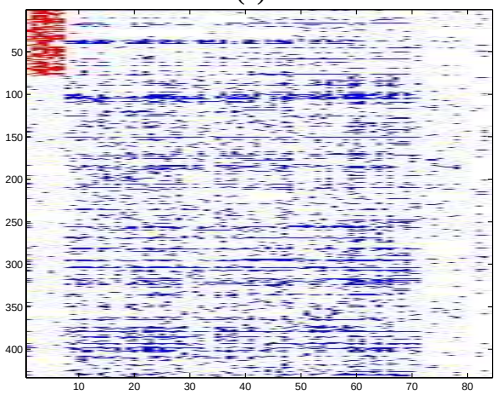
[4] S. Lonardi, W. Szpankowski, and Q. Yang. Finding biclusters by random projections. In *Proc. 15th Annual Combinatorial Pattern Matching Symp. (CPM'04)*, 2004.

[5] I. Shmulevich and W. Zhang. Binary analysis and optimization-based normalization of gene expression data. *Bioinformatics*, 18:555–565, 2002.

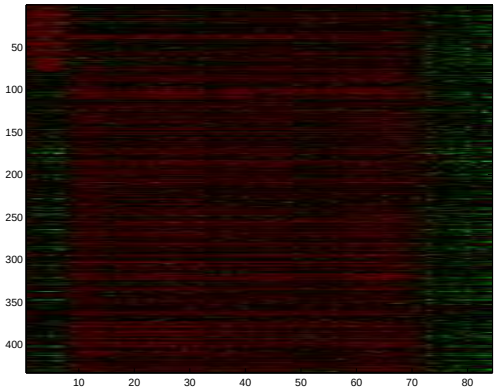
[6] A. Tanay, R. Sharan, and R. Shamir. Discovering statistically signifi cant biclusters in gene expression data. *Bioinformatics*, 18:S136–S144, 2002.




(a)




(b)




(c)

**Figure 4. A locally maximal significant bi-cluster discovered in breast tumor data. (a) Bicluster in the binary quantized matrix, (b) binary matrix reordered to bring rows and columns in bicluster together, (c) original gene expression matrix reordered accordingly.**