Communications

Co-clustering: A Versatile Tool for Data Analysis in Biomedical Informatics

Sungroh Yoon, Luca Benini, and Giovanni De Micheli

***Abstract*—Co-clustering has not been much exploited in biomedical informatics, despite its success in other domains. Most of the previous applications were limited to analyzing gene expression data. We performed co-clustering analysis on other types of data and obtained promising results, as summarized in this paper.**

***Index Terms*—Acute myeloid leukemia, biomedical informatics, co-clustering, microRNA, single nucleotide polymorphism.**

I. Introduction

Clustering may be the most popular technique for data analysis in many disciplines. Unlike conventional clustering that groups similar object from a single collection of objects, coclustering or biclustering [1] aims to group objects from two disjoint sets simultaneously, thus, revealing interactions among the elements in the two sets. A well-known example occurs in text mining, where documents are grouped based on their word contents. Although coclustering can be a powerful tool for analyzing various biomedical data, the applications of coclustering in the community have been focused mostly on gene expression analysis [2], [3]. In this paper, we report several applications of coclustering to large-scale biomedical data sets other than gene expression data.

The problem of coclustering can be formulated as that of finding complete subgraphs (bicliques) in a bipartite graph. (For computational efficiency or for robustness, the completeness of subgraphs can be relaxed, and some methods search "approximate" bicliques instead). Most of the coclustering techniques assume a weighted bipartite graph, and find bicliques with some conditions imposed on the weights of edges in bicliques. Alternatively, if the input is given as a matrix, coclusters can be regarded as (possibly overlapping) submatrices in which (some or all) elements satisfy some specific conditions (e.g., the values on each row are similar).

It is known that the problem of coclustering is NP-hard [1], and many techniques have been proposed to cope with this computational challenge. A review of coclustering algorithms can be found in [1], and a performance comparison among some coclustering algorithms is presented in [2], [4].

Manuscript received July 15, 2006. This work was supported in part by a grant from Jerry Yang and Akiko Yamazaki.

S. Yoon was with the Computer Systems Laboratory, Stanford University, Stanford, CA 94305 USA. He is now with Intel Corporation, Santa Clara, CA 95054 USA (e-mail: sryoon@gmail.com).

L. Benini is with the Department of Electrical Engineering and Computer Science, University of Bologna, 40136 Bologna, Italy (e-mail: lbenini@deis.unibo.it).

G. De Micheli is with the Integrated Systems Center, Ecole Polytechnique Federale de Lausanne, CH-1015 Lausanne, Switzerland (e-mail: giovanni.demicheli@epfl.ch).

Digital Object Identifier 10.1109/TITB.2006.897575

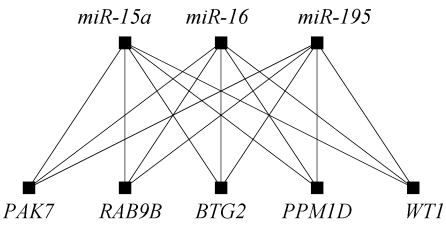

Fig. 1. Cocluster of microRNAs and target genes found in the human genome [6]. Genes *BTG2*, *WT1*, *PPM1D*, *PAK7*, and *RAB9B* are mostly regulators, and their anomaly can be found in breast, renal, and prostate cancers. The miRNAs *miR-15a* and *miR-16* are clustered on chromosome 13q14, and this region has been shown to be deleted altogether in several types of cancer.

II. Applications of Coclustering

A. Predicting MicroRNA Regulatory Modules

One of the most important advances in biology in recent years may be the discovery of RNAs that can regulate gene expression. As one kind of such functional noncoding RNAs, microRNAs (miRNAs) form a class of endogenous RNAs that can have important regulatory roles by targeting transcripts for cleavage or translational repression [5].

Multiple miRNAs often regulate a common mRNA whereas one miRNA may have several target genes. This multiplicity of targets and cooperative signal integration on target genes are key features of the control of translation by miRNAs [5], [6] but this also makes it challenging to detect important patterns appearing in the gene regulation mechanism by miRNAs.

To address this issue, we developed a coclustering-based method that can computationally predict *miRNA regulatory modules* (MRMs) or coclusters of miRNAs and their targets that are believed to participate cooperatively in post-transcriptional gene regulation [6]. This method was tested with the human genome, and approximately 400 MRMs were identified. Fig. 1 shows an MRM discovered from this study.

B. Linking Clinical Traits With Gene Expression

The invention of DNA microarray technologies has enabled researchers to simultaneously monitor the expression level of a whole genome. Thus, for the purpose of finding genes related to a certain clinical trait of interest, it has become feasible to examine all the genes available and then select only those whose expression is consistently correlated with the trait over many samples. Although correlation does not always imply causality, this approach has been successful in many studies as an attempt to understand genetic mechanisms underlying clinical observations [7]–[10].

Example 1: It is possible to calculate the correlation coefficient between one row vector of the trait matrix in Fig. 2(a) and another row vector of the expression matrix in Fig. 2(c). For example, the points on the red curve in Fig. 2(d) represent the Pearson correlation coefficients between trait 1 and the genes in the expression matrix. By inspection, we can observe that genes 2, 3, and 4 are correlated with traits 1 and 2.

Obviously, this inspection method breaks down as the size of a problem grows, and clearly there is a need for a computational method that can automatically detect patterns appearing on correlation curves. Using coclustering, we introduced a method that can automatically reveal complex relationships between multiple genes and traits [11]. This technique finds coclusters of genes and clinical traits and was tested

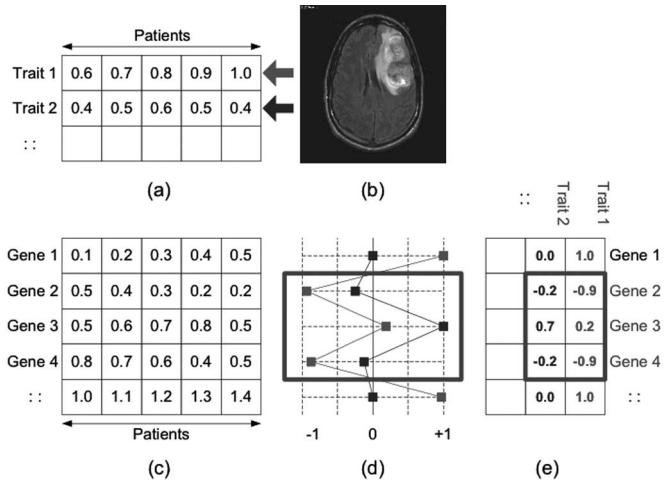

Fig. 2. Example of coclustering genes and clinical traits (courtesy of Dr. Michael D. Kuo at UCSD; the presented numbers replaced with fictitious values for confidentiality). (a) Matrix of clinical traits (e.g., tumor size) derived from the image in (b). (b) Brain image. (c) Gene expression matrix. Columns are arranged in the same order as in (a). (d) Plot showing the correlation coefficient between two rows in the matrices in (a) and (b). (e) Matrix representation of the plot in (c). Coclusters of genes and clinical traits can be found here.

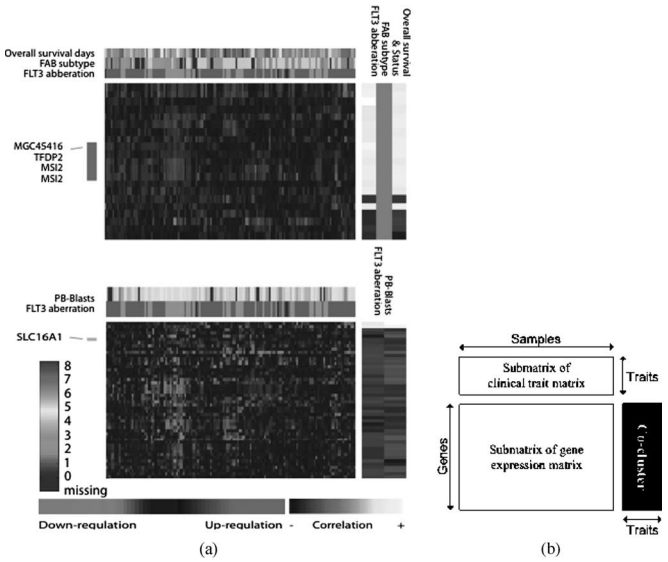

Fig. 3. (a) Heat-map representations of two coclusters found from a public AML data set by the technique proposed in [11]. (b) Each representation consists of three panels, and a cocluster corresponds to the panel in the right.

with a public acute myeloid leukemia (AML) data set [10], discovering 43 statistically significant coclusters. Two of these coclusters are shown in Fig. 3.

C. Single Nucleotide Polymorphism Studies

Genome-wide association studies is to obtain information on the association of *single nucleotide polymorphism* (SNP) to phenotypes across the entire genome. SNPs are the most common form of genetic variation in humans comprising almost 0.1% of the average human genome. Predicting and understanding the downstream effects of genetic variation using computational methods are becoming increasingly important for SNP selection in genetic studies [12].

Most SNP data sets consists of a subject-by-genotype matrix in which rows correspond to subjects under study and columns to SNP genotypes for each subject, as well as a subject-by-phenotype matrix that records phenotypes for each subject. By a correlation analysis similar to that mentioned in Section II-B, it is possible to find a genotype-by-phenotype matrix. Finding coclusters appearing in this matrix corresponds to associating genotypes with phenotypes, and evaluation of these coclusters may provide valuable biological insight.

III. Conclusion

The task of examining objects in two disjoint sets and revealing implied interactions occur frequently in biomedical informatics, and coclustering can be a versatile and powerful data-mining tool to complete this task.

References

[1] S. C. Madeira and A. L. Oliveira, "Biclustering algorithms for biological data analysis: A survey," *IEEE/ACM Trans. Comput. Biol. Bioinformat.*, vol. 1, no. 1, pp. 24–45, Jan./Mar. 2004.

[2] A. Tanay, R. Sharan, and R. Shamir, "Discovering statistically significant biclusters in gene expression data," *Bioinformatics*, vol. 18, pp. S136–S144, 2002.

[3] Y. Cheng and G. M. Church, "Biclustering of expression data," in *Proc. ISMB*, 2000, pp. 93–103.

[4] S. Yoon, C. Nardini, L. Benini, and G. De Micheli, "Discovering coherent biclusters from gene expression data using zero-suppressed binary decision diagrams," *IEEE/ACM Trans. Comput. Biol. Bioinformat.*, vol. 2, no. 4, pp. 339–354, Oct./Dec. 2005.

[5] D. P. Bartel, "MicroRNAs: Genomics, biogenesis, mechanism, and function," *Cell*, vol. 116, no. 2, pp. 281–297, 2004.

[6] S. Yoon and G. De Micheli, "Prediction of regulatory modules comprising microRNAs and target genes," *Bioinformatics*, vol. 21, pp. ii93–ii100, Sep. 2005.

[7] A. R. Whitney, M. Diehn, S. J. Popper, A. A. Alizadeh, J. C. Boldrick, D. A. Relman, and P. O. Brown, "Individuality and variation in gene expression patterns in human blood," *Proc. Nat. Acad. Sci. USA*, vol. 100, no. 4, pp. 1896–1901, Feb. 2003.

[8] M. West, C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, J. A. Olson Jr, J. R. Marks, and J. R. Nevins, "Predicting the clinical status of human breast cancer by using gene expression profiles," *Proc. Nat. Acad. Sci. USA*, vol. 98, no. 20, pp. 11462–11467, Sep. 2001.

[9] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proc. Nat. Acad. Sci. USA*, vol. 98, no. 9, pp. 5116–5121, Apr. 2001.

[10] L. Bullinger, K. Dohner, E. Bair, S. Frohling, R. F. Schlenk, R. Tibshirani, H. Dohner, and J. R. Pollack, "Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia," *N. Eng. J. Med.*, vol. 350, no. 16, pp. 1605–1616, Apr. 2004.

[11] S. Yoon, L. Benini, and G. De Micheli, "Finding co-clusters of genes and clinical parameters," in *Proc. 27th Annu. Int. Conf. IEEE EMBS*, Sep. 2005, pp. 906–912.

[12] S. Mooney, "Bioinformatics approaches and resources for single nucleotide polymorphism functional analysis," *Brief. Bioinformat.*, vol. 6, no. 1, pp. 44–56, Mar. 2005.