

A Bi-ordering Approach to Linking Gene Expressions with Clinical Annotations in Cancer

Fan Shi¹, Geoff MacIntyre¹, Chris Leckie¹, Izhak Haviv², Alex Boussioustas³, Adam Kowalczyk¹
{shif,gmaci,caleckie}@csse.unimelb.edu.au, Izhak.haviv@bakeridi.edu.au, akowalczyk@nicta.com.au

¹ National ICT Australia & The University of Melbourne, Parkville, Victoria 3010, Australia;

² Baker IDI, 75 Commercial Road Prahran, Victoria. 3004, Australia;

³ Peter MacCallum Cancer Centre, St. Andrew's Place, East Melbourne, Victoria. 3002, Australia.

Abstract

In this paper we introduce a robust method for exploratory biclustering analysis of microarray data, which produces a number of different bi-orderings of the data, each uniquely determined by a bicluster, i.e., a pair of subsets of genes and samples. The core algorithm is closely related to biclustering. We show that the biclusters and bi-orderings generated by our method are highly statistically significant with respect to both the sample histological annotations and biological annotations (from the Gene Ontology). Some of the gene modules associated with our most robust bi-orderings are closely linked to gene modules that are important for gastric cancer tumor genesis reported in literature, while others are novel discoveries.

Keywords: biclustering, gene expression, gastric cancer, gene ontology

1 Introduction

A typical aim of exploratory analysis of genomics data is to identify potentially interesting genes or pathways that warrant further investigation. There is a critical need to streamline this type of analysis, in order to support continuing advances in high throughput genomics methods such as gene expression microarrays, which measure thousands of genes in a single assay and are the focus of this paper. Such assays provide noisy and incomplete measurements, which require sophisticated bioinformatics techniques to identify statistically and biologically significant associations between genes and relevant phenotypes of interest.

This paper proposes such an exploratory technique. It is based on a form of biclustering [5], i.e., a method for automated discovery of highly correlated subsets of genes across a subset of samples, in combination with methods for evaluating the statistical significance and biological relevance of such biclusters. There are four main contributions that we make in this paper. First, we introduce a novel algorithm, called bi-ordering, which is in some respects a member of a family of biclustering techniques [2][3][4][6]. Second, we introduce two novel statistical techniques for evaluating the significance of the generated groupings and orderings of multiple histological samples. Third, we assess the stability of the observed results by assessing the size of their “basin of attraction” as follows. In our experiments, random initializations of the algorithms yield hundreds of biclusters, which were then grouped into a manageable number of families of identical or very similar outcomes (called “super-biclusters”) by a secondary phase of clustering the generated biclusters. The size of such a family is interpreted as the stability of the super-bicluster. We found that our technique can find a small set of highly stable super-biclusters, which correspond to distinct histopathological types in an existing gastric cancer data set [1]. Fourth, we demonstrate that the discovered super-biclusters have associated Gene Ontology (GO) terms with very significant p-values, which can serve as a basis for biological interpretation of the meaning of the associated gene modules.

2 Method and Results

We introduce a protocol for identifying and characterizing modules of genes that exhibit high statistical, biological and clinical significance. Our protocol, named Bi-ordering Exploratory Analysis (BEA), comprises six main stages as stated below:

1. Input: $n_G \times n_S$ gene expression data matrix of n_G genes for n_S samples
2. Generate biclusters based on a bi-ordering of genes and samples, called bi-ordering analysis (BOA)
3. Merge similar biclusters into “super-biclusters” to identify robust modules of co-expressed genes
4. Annotate biclusters with histological and biological attributes to support their interpretation
5. Generate figures of merit (i.e., p-values) for:
 - a. GO annotations;
 - b. saturation metric: overrepresentation of histological categories in bicluster, and;
 - c. trend statistics: concordance of sample order with various phenotype gradients
6. Develop biological interpretation of the results.

In this paper, we analyze the performance of our algorithm on a real gene expression dataset for gastric cancer [1]. The main reason for this choice is the availability of local expertise in the biology of this disease. Gastric cancer dataset contains 7383 genes and 124 samples from 6 pathological categories (CG, IM and normal are subtypes of premalignant samples; Diffuse, Intestinal and Mixed are malignant samples). We evaluate the results using the three metrics mentioned in the previous section. By applying the BOA algorithm on gastric cancer dataset, a group of biclusters are generated. An example of the heat map of a bicluster is shown in Fig. 1, and the p-values in terms of different metrics are shown in Table 1.

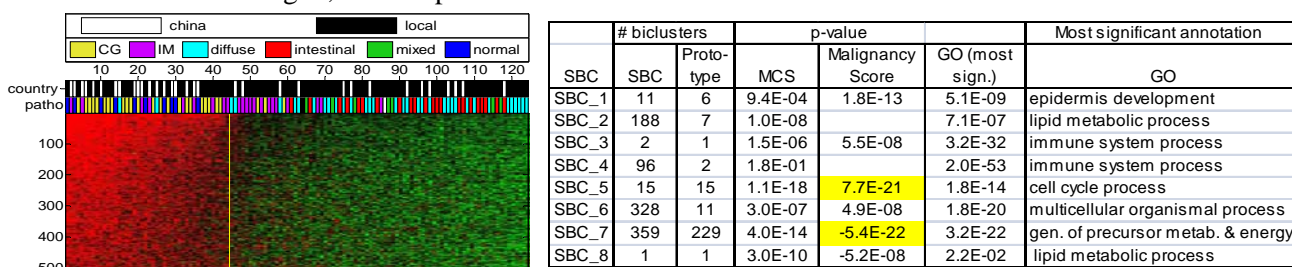


Fig. 1. Heat map for the most prominent super-bicluster, SBC_7, generated by the BOA algorithm for the gastric cancer data. We observe the strong gradation from least “malignant” normal samples, though CG and IM, to the malignant samples (combined intestinal, diffuse and mixed gastric cancers). The vertical axis shows the 515 most significant genes, while the horizontal axis shows the final order of samples generated by the BOA algorithm. The yellow vertical line indicates the right boundary of samples in the bicluster.

Table 1. Numerical characterisation and biological relevance of super biclusters generated by BOA. Note that the negative sign, ‘-’, in Malignancy Score for SBC_7 and SBC_8 indicates the significance of agreement with the reverse order. In the second and third columns of the table, the number of biclusters that converged to a particular super-bicluster or its prototype are given.

The generated results including the GO annotations and clinical correlations were the basis of an evaluation by expert biologists and clinicians who judged that our BEA protocol has generated a number of significant biological hypotheses warranting follow-up wet lab experiments. In summary, the BOA results have shed new light on preexisting themes in gastric cancer etiology. The resulting bi-orderings represent successive steps in cancer progression or distinct correlations with pathological types of the disease.

References

- [1] A. Boussioutas, et al., *Cancer Research*, 63, pp. 2569–2577, 2003.
- [2] A. Tanay, R. Sharan and R. Shamir, *Bioinformatics*, Vol. 18, pp. S136-S144, 2002.
- [3] J. Ihmels, S. Bergmann and N. Barkai, *Bioinformatics*, Vol. 20, No. 13, pp. 1993-2003, 2004.
- [4] Q. Sheng, Y. Moreau, and B. De Moor, *Bioinformatics*, Vol. 19, pp. ii196-ii205, 2003.
- [5] S. C. Madeira, A. L. Oliveira, *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, pp.196-205, 2004.
- [6] Y. Cheng and G. M. Church, *ISMB*, 2000.