

Formalizing a New Model-based Method of Classifying Gene Expressions

Vinhthuy Phan,¹ Raghuver Kontham,² E. Olusegun George³ Tom Sutter⁴

Keywords: gene expression, clustering, multiple-comparison hypothesis test

1 Classifying Gene Expression Using Hypothesis Tests

Classifying genes in a biologically meaningful way is a challenging and important problem that biologists face today. Microarray technology has been very helpful in differentiating the expression of genes under different conditions. And many methodologies have been developed to infer biological meanings by grouping similarly expressed genes. Nevertheless, there has been no method to date that is universally accepted as the best method of classifying gene expressions. Sutter et al. [2] developed an innovative way to classify genes, based on outcomes of statistical experiments. The method uses the results of non-parametric multiple comparisons, which resulted in the bipolar hierarchical clustering of genes in relation to their response to treatment. Essentially, each gene is placed in a cluster, which corresponds to its pattern of outcomes from all possible two-sided hypothesis tests among all pairs of treatments. These outcomes, though based entirely on statistical construction, reveal meaningful biological information about the clusters that the genes fall into. It was shown that this method reveals more biological information than do other popular classifying methods such as hierarchical clustering and principle component analysis [1].

As this method relies on analyzing the outcomes of multiple-comparison hypothesis tests, it is important to understand the patterns of these outcomes in terms of characterizing and generating them. For experiments involving gene responses to different treatments, an enumeration of all possible *meaningful* patterns of outcomes, along with their biological characteristics, is more desirable before the tests are carried out, as this could be used to focus on more interesting clusters of genes before any hypothesis is tested. In this article, we formalize the properties and patterns of the outcomes obtained from pairwise two-sided hypothesis tests for any given number of treatments. We study the problem of counting and enumerating the number of such patterns, and of identifying "interesting" clusters of patterns. While counting all possible patterns is an open problem, we are able to count a smaller and more meaningful subset of all patterns of outcomes. Further, we demonstrate that biologically "interesting" clusters of genes do correspond to some of the "interesting" clusters of genes that we formalize.

2 Characterizing Outcomes of Multiple Pairwise Comparisons

Unlike methods like hierarchical clustering, where genes having (quantitatively) similar levels of expression, are grouped into common clusters, the method proposed by Sutter et al. [2] attempts to classify gene expressions in a more elaborate fashion. To classify gene responses to n different treatments, Sutter et al. measure the statistically significant, *relative* difference

¹Dept of Mathematical Sciences, The University of Memphis, TN. E-mail: vphan@memphis.edu

²Dept of Mathematical Sciences, The University of Memphis, TN. E-mail: rkontham@memphis.edu

³Dept of Mathematical Sciences, The University of Memphis, TN. E-mail: eogeorge@memphis.edu

⁴Dept of Biology, The University of Memphis, TN. E-mail: tsutter@memphis.edu

of expression for any two genes when they are under influence of any two of the n treatments. This way, every pattern of outcome out of $\binom{n}{2}$ possibilities corresponds to exactly one group of genes. For a two-sided hypothesis test, there are $3^{\binom{n}{2}}$ possible patterns, not all of which can be observed. Those that are meaningful are still numerous to reveal interesting information before the hypothesis tests are carried out. It is, therefore, important to understand as much about these patterns as possible. Our objective is to characterize and enumerate them. Additionally, when the tests are finished, and genes are labelled with corresponding patterns, each representing the pairwise responses of those genes to n treatments, we would like to classify these patterns further to provide a more complete picture about the relationship of the gene responses. Denote $x < y$ as the outcome of a hypothesis test that results in a gene responding more to treatment y in comparison to its response to treatment x ; and denote $x \sim y$ as the outcome of a hypothesis test that results in a gene whose response to x and y is not distinguishable; statistically, the null hypothesis $x = y$ is accepted. Note that, given 3 treatments for instance, while it is possible to observe $x \sim y, x < z, y < z$, it is not possible to observe $x < y, y < z, z < x$.

Sutter et. al. proposed a tree representation of the patterns of outcomes that is good for a simple visualization. In this work, we offer two other representations: partial order sets and graphs. As partial order sets, we observe that currently there is no formula for the number of possible meaningful outcomes when $n > 13$. This representation allows us to count approximately the number of such outcomes by restricting the set of meaningful outcomes to only cases where treatment x and y are identical if there is no statistical difference in their gene responses. As a result, the treatments can be grouped into statistically distinguishable, equivalence classes. This subset consists of only totally ordered sets, and as such we can enumerate all meaningful representations. $\sum_{k=1}^n \left\{ \begin{matrix} n \\ k \end{matrix} \right\} k! = \sum_{k=1}^n \sum_{i=0}^{k-1} (-1)^i \binom{k}{i} (k-i)^n$ where $\left\{ \begin{matrix} n \\ k \end{matrix} \right\}$, known as Sterling's number of the second kind, is the number of ways of partitioning a set of n treatments into k non-empty parts.

After placing genes into each cluster represented by a pattern, it is desirable to further cluster these groups of clusters to give a finer view on the relationship among these genes. The graph representation, where vertices represent treatments and edges represent the relative response of a gene in two different treatments (either $x < y$ or $x > y$), helps us to achieve this purpose. The post-classification of patterns employs the hierarchical clustering and the k-mean clustering algorithms. Using hierarchical clustering, we define a modified Hamming metric on the space of patterns. Using k-mean, we define the initial k centroids as the k patterns whose gene responses are most *distinctive*. For example, the pattern representing the cases where all genes response positively to a certain treatment would be viewed as the most distinctive pattern. The degree of distinctiveness of a pattern can be mathematically defined in terms of the difference between the in- and out- degrees of its graph representation together with the number of genes falling into that pattern.

References

- [1] Raychaudhuri, S., Stuart, J.M., and Altman, R.B., "Principle components analysis to summarize microarray experiments: application to sporulation time series", , *Pac. Symp. Biocomput.*, 2000: pp. 455-466, 2000.
- [2] Sutter, T., He, X.R., et. al, "Multiple Comparisons Model-based Clustering and Ternary Pattern Tree Numerical Display of Gene Response to Treatment: Procedure and Application to the Preclinical Evaluation of Chemopreventive Agents", *Molecular Cancer Therapeutics*, Vol. 1, pp. 1283-1292, Dec. 2002.