論文の内容の要旨

応用生命工学専攻
平成２８年度博士課程　入学
氏名　趙　世涛
指導教員　清水　謙多郎

論文題目

Application of Silhouette Scores for Arbitrarily Defined Groups in Gene Expression Data
（遺伝子発現データにおける任意定義群に対するシルエットスコアの応用）

Introduction

RNA plays an important role in the central dogma of molecular biology as a key intermediate between a genome and proteome. Quantifying the expression levels of transcripts under different conditions is a basic task in transcriptomics analyses, which include two key contemporary techniques, namely microarrays and RNA sequencing (RNA-seq). These two methods are ideal for the analysis of differential gene expression (DGE) and have become routine tools in the fields of molecular biology, medicine, agriculture, and ecology.

Hierarchical sample clustering (HSC) is performed widely to examine associations between expression data obtained from microarrays and RNA-seq based on similarities in expression patterns. Some researchers empirically know that an HSC result of data designed for DGE analysis roughly corresponds to the DGE result when the groups for the DGE analysis are evaluated with respect to the HSC result. If individual groups form distinct subclusters consisting only of members (or samples) from a particular group, DGE analysis performed using such distinct groups would result in many differentially expressed genes (DEGs). Conversely, if the members (or samples) of each subcluster originate from multiple groups, no or few DEGs would be expected. However, an objective evaluation of the relationship between the results of HSC and the percentage of DEGs (P_{DEG}) value remains to be performed. Numerical scores that indicate the degree of separation between predefined groups would benefit the objective assessment of the results of HSC.

In this study, I propose the use of a silhouette score for the objective evaluation of gene expression data based on arbitrary grouping criteria. I evaluated comprehensively the relationship among the results of HSC, DGE, and silhouette score using both simulated and real expression data obtained from RNA-seq and microarrays.

Methods

A rough correlation between the shape of the HSC dendrogram and P_{DEG} in gene expression data has been reported (Tang et al., 2015). The two-group comparison of gene expression data has been used as an example to explain this correlation. Using HSC, the generation of two distinct clusters (Cluster1 and

Cluster2), each including members that are consistent with known group information (Group1 and Group2), tends to yield a larger P_{DEG} value in the DGE analysis. Conversely, if HSC produces clusters that seem to be intermixed and cannot be completely divided into distinct ones, the P_{DEG} value becomes smaller and infinitely close to zero; researchers in related fields have always empirically considered that to be a reasonable explanation. Because the interpretation of the results of HSC is still used in subjective visual inspection, herein, I attempted to evaluate these results using an objective score and group label information.

In this study, I proposed the use of silhouette score as an interpretation criterion for objectively evaluating the degree of separation between two or more groups under comparison (Zhao et al., 2018). Although silhouette is generally used for the validation of clustering results, I here employed it independently from clustering. Technically, the term cluster is replaced with group in the silhouette calculation procedure. A silhouette score $s(i)$ obtained for each sample i is calculated using the following equation:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

where $a(i)$ is the mean distance between i and all other samples in the same group (e.g., group A) and $b(i)$ is the smallest mean distance of i to all samples in any other group, of which i is not a member. The $s(i)$ ranges from − 1 to 1; it is positive if $a(i) < b(i)$, zero if $a(i) = b(i)$, and negative if $a(i) > b(i)$. A larger $s(i)$ value indicates increased group separation and vice versa. By taking the mean $s(i)$ over all samples, the average silhouette (AS) value for each comparison can be obtained. To the best of my knowledge, the current study is the first practical application of the concept to estimate the degree of separation between groups (not clusters) using gene expression data.

Results and discussions

As an example of the results of such analysis, the results of HSC of a three-group RNA-seq data (Blekhman et al., 2010) obtained for three species (i.e., humans (HS), chimpanzees (PT), and rhesus macaques (RM)) are shown in Figure 1. Briefly, Blekhman et al. studied gene expression levels in liver samples from three males (M1, M2, and M3) and three females (F1, F2, and F3) from each species/group. Henceforth, I will use the corresponding abbreviations (HSF, HSM, PTF, PTM, RMF, and RMM) for better readability. The P_{DEG} values satisfying the 10% false discovery rate threshold were obtained using TCC package (Sun et al., 2013) in R, with default settings.

Because the branch lengths in the HSC dendrogram corresponding to the vertical axis in Figure 1 were calculated using "1 - Spearman's rank correlation coefficient (r)," a shorter branch length implies a closer expression pattern between the samples and vice versa. It can be seen that the HSC dendrogram is clearly divided into three clusters among the three species (HS, PT, and RM) and the expression pattern of two groups, HS and PT, is similar against that of HS and RM. Additionally, in the same cluster (leftmost

of the three largest clusters), female (HSF) and male (HSM) samples are intermixed.

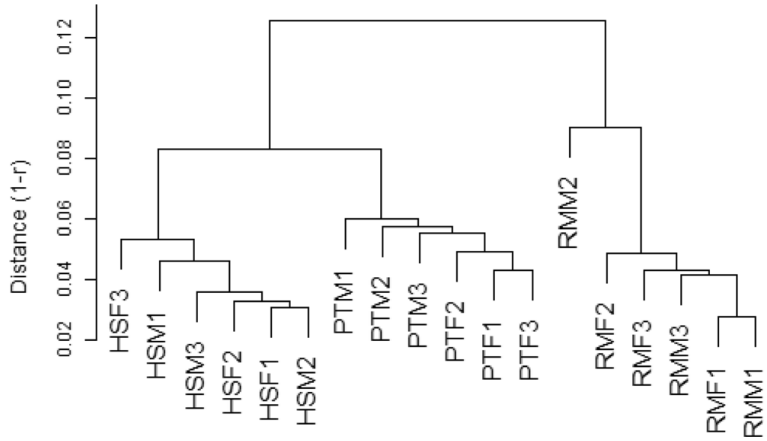

Figure 1 Results of HSC obtained using Blekhman's data.

Table 1 shows the partial results of the two-group comparisons (P_{DEG} values and corresponding AS values). A greater AS value yielded a higher P_{DEG} value, whereas a smaller AS value resulted in a lower P_{DEG} value. This is consistent with the lack of separation between female and male samples within each species in the HSC dendrogram.

Table 1 Results of the comparison of two arbitrarily defined groups from Blekhman's data

	P_{DEG} (%)	AS
HSF vs HSM	0.07	−0.019
HSF vs PTF	7.56	0.389
HSF vs RMF	19.75	0.590

Based on our visual evaluation, the AS values effectively represented the overall relationship between groups of interest in the HSC analysis. I think the expressive power in cases of few or no DEGs in the dataset (i.e., AS ≈ 0) is practically promising, but increasing the correlation between P_{DEG} and AS is not practical. This is simply because the P_{DEG} value tends to increase as the number of replicates (N_{rep}) increases, suggesting that the correlation is influenced by N_{rep}. We next investigated the effects of N_{rep} on P_{DEG} and AS values. As a result, I found that AS values were independent of N_{rep}, while P_{DEG} values obtained from DGE analysis were fundamentally dependent on N_{rep}.

To further investigate the relationship between P_{DEG} and AS values under a fixed N_{rep} value, I performed two-group comparison using both simulated and real RNA-seq data. The two-group simulated data were produced using the “simulateReadCounts” function in TCC. Three real RNA-seq datasets were used in the analysis; i.e., a yeast dataset (Schurch et al., 2016), a mouse dataset (Bottomly et al., 2011), and a human dataset (Cheung et al., 2010). All the results support the conclusion that was drawn from the aforementioned example.

I also investigated two microarray datasets (Nakai et al., 2008; Kamei et al., 2013). I observed similar results between the Nakai and Blekhman's data and between the Kamei and Cheung's dataset. In particular, the P_{DEG} and AS values obtained using the latter datasets were close to zero and the HSC dendrogram showed an intermixed structure. These results indicate that AS can be utilized as supporting information to interpret the results of DGE for both RNA-seq and microarray data, especially when no or few DEGs are obtained.

Conclusion

I proposed the use of silhouette score (i.e., AS values) as an objective measure of the degrees of separation between groups of interest based on expression data. Silhouettes is useful for exploring data with predefined group labels. It would help provide both an objective evaluation of HSC dendrograms and insights into the DGE results with regard to the compared groups. The use of this measure would enable a more objective discussion about the HSC result in terms of the groups.

References

Blekhman R, Marioni JC, Zumbo P, Stephens M, et al., Sex-specific and lineage-specific alternative splicing in primates. *Genome Res*., **20**: 180-189, 2010.

Bottomly D, Walter NA, Hunter JE, Darakjian P, et al., Evaluating gene expression in C57BL/6J and DBA/2J mouse striatum using RNA-Seq and microarrays. *PLoS One*, **6**: e17820, 2011.

Cheung VG, Nayak RR, Wang IX, Elwyn S, et al., Polymorphic cis- and trans-regulation of human gene expression. *PLoS Biol*., **8**: e1000480, 2010.

Kamei A, Watanabe Y, Kondo K, Okada S, et al., Influence of a short-term iron-deficient diet on hepatic gene expression profiles in rats. *PLoS One*, **8**: e65732, 2013.

Nakai Y, Hashida H, Kadota K, Minami M, et al., Up-regulation of genes related to the ubiquitin-proteasome system in the brown adipose tissue of 24-h-fasted rats. *Biosci Biotechnol Biochem.* **72**: 139-148, 2008.

Schurch NJ, Schofield P, Gierliński M, Cole C, et al., How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA*, **22**: 839-851, 2016.

Sun J, Nishiyama T, Shimizu K, Kadota K. TCC: an R package for comparing tag count data with robust normalization strategies. *BMC Bioinformatics*, **14**: 219, 2013.

Tang M., Sun J, Shimizu K, Kadota K. Evaluation of methods for differential expression analysis on multi-group RNA-seq count data. *BMC Bioinformatics*, **16**: 360, 2015.

Zhao S, Sun J, Shimizu K, Kadota K. Silhouette Scores for Arbitrary Defined Groups in Gene Expression Data and Insights into Differential Expression Results. *Biol Proced Online*, **20**: 5, 2018.