Development of high-accuracy clustering algorithm based on statistical test results

Aoto Yoshimasa[1,a)] Hachiya Tsuyoshi[2] Okumura Kazuhiro[3] Hase Sumitaka[1] Sato Kengo[1]
Wakabayashi Yuichi[3] Sakakibara Yasubumi[1]

Abstract: In recent years, RNA-Seq technologies have been applied efficiently for comprehensive detection and quantification of genes expressed in cells. Genes are stochastically expressed with large variances; therefore, it is important to discriminate significant differences from insignificant differences, such as noises between expression levels. However, it is difficult to assess stochastic dispersion in data using the existing definitions of distance for clustering. In this paper, we propose a novel gene clustering algorithm, digital clustering, for the analysis of multi-conditional transcriptome dynamics. For digital clustering, a novel definition of distance between clusters was adopted and a hierarchical method was applied. Specifically, distance is based on the results of statistical tests for all pairs of conditions for each gene; this enables digital clustering to discriminate significant differences from insignificant differences in expression levels. From our benchmark, our algorithm achieved high classification accuracy for multi-conditional transcriptome data.

background

In the last decade, high-throughput sequencing technologies have been widely used for research in genomics, transcriptomics, and other areas. Especially, transcriptome analyses are indispensable for measuring gene expression and direct comparisons of gene expression levels between case and control samples to detect differentially expressed genes (DEGs). Recently, RNA sequencing (RNA-Seq) technology has been used to comprehensively detect and quantify cellular gene expression [1, 2].

After quantifying gene expressions by RNA-Seq, genes are often clustered according to expression pattern to predict gene functions, search for biomarkers, and summarize large data sets. For clustering by gene expression, Euclidean distance, Pearson's correlation, and cosine distance are commonly used as measures of distance or similarity between genes [3]. Since gene expression is a stochastic process [4], there are both significant differences and insignificant differences between expression levels. Therefore, it is important to discriminate significant expression differences from noise, but this is impossible with the existing definitions of distance, which do not account for dispersion in data.

In this paper, we propose a novel gene clustering algorithm, digital clustering, for the analysis of multi-conditional transcriptome dynamics. Digital clustering is a hierarchical method. It adopts a novel definition of distance that is based on the results of statistical tests for all pairs of conditions for each gene. This not only enables digital clustering to be used to distinguish random from nonrandom differences in expression levels, but also facilitates its application to multi-conditional (e.g., time-course) transcriptome data.

[1] Keio University, 3-13-1 Hiyoshi, 223-8522 Yokohama, Japan
[2] Iwate Medical University, 2-1-1 Nishitokuta, Yahaba-cho, 028-3694 Shiwa-gun, Japan
[3] Chiba Cancer Center Research Institute, 666-2 Nitonacho, Chuo Ward, Chiba, 260-*8717 Chiba, Japan
[a)] aoyocchi@dna.bio.keio.ac.jp

Methods

Digital clustering

Digital clustering is a novel hierarchical clustering algorithm for the analysis of multi-conditional (e.g., time-course) transcriptome data. First, an integer vector, called the *digital vector*, is assigned to each gene as follows. In accordance with common analytical pipelines, short reads from RNA-Seq are mapped to a reference genome, and normalized expression values are calculated. Using these normalized values, a statistical test is applied to all pairs of conditions (samples) to detect DEGs. On the basis of statistical test results, we assign one of three values to each pair; +1 for up-regulated pairs, −1 for down-regulated pairs, and 0 for insignificant pairs. In this digital vector, each element corresponds to the statistical test result for the comparison between conditions (samples). The expression level of a gene g in condition k is denoted e_k^g. If the expression level e_k^g is significantly different from e_l^g, then the element of the corresponding dimension in the digital vector is $1(e_k^g < e_l^g)$ or $-1(e_k^g > e_l^g)$; otherwise, it is 0.

Second, a cluster of genes is associated with a digital vectors as follows. The element that corresponds to the statistical test result for condition k and condition l in the digital vector of cluster C_n is 1 if $d_{k,l}^{\forall g \in C_n} = 1$ and -1 if $d_{k,l}^{\forall g \in C_n} = -1$; otherwise, it is 0.

Digital clustering is a hierarchical method for clustering genes by using a digital vector as described above. The distance between two clusters C_n and C_m is defined as follows:

$$D(C_n, C_m) = 1 - \frac{(\text{gain}_{n,m} - \text{loss}_{n,m})}{(\text{gain}_{n,m} + \text{loss}_{n,m})}. \quad (1)$$

Let s denote the number of elements of 1 or -1 in the digital vector $v(C_n \cup C_m)$ associated with the union of clusters C_n and C_m. Then, $\text{gain}_{n,m}$ is defined as $s \times |C_n \cup C_m|$. Let t denote the total number of elements of 1 or -1 in the digital vectors $v(g)$ with all g in the union. The $\text{loss}_{n,m}$ is defined to be $(t - \text{gain}_{n,m})$. Note that gains and losses are non-negative integer values. If $(\text{gain}_{n,m} + \text{loss}_{n,m}) = 0$, then $D(C_n, C_m) = 0$. If any pairs of clusters have the same distance under the above definition, then the digital clustering compares Pearson's correlation of expression values among them using the group average method.

Digital clustering iteratively searches for and merges the pair with minimum distance, constructing hierarchical trees based on statistical test results. Initially, genes that have the same statistical test results are merged because the distance between genes with the same digital vector is zero. Subsequently, digital clustering merges clusters so as to conserve common test results as much as possible.

Simulation analysis

For benchmark testing, we generated simulated short reads from mouse reference cDNA sequences so that gene expression profiles for four artificial conditions were obtained. We set up the ten digital vectors as the correct DEG labels, and 20% or 10% of genes were randomly selected as DEGs between pairs of conditions so that each DEG belonged to any of ten clusters. The read counts were generated from negative binomial distributions with the mean and the variance which were extracted from Pickrell et al. data [5]. According to the analytical protocol of Trapnell et al. [6], we estimated gene expression levels and searched for DEGs from generated reads. The statistical test results were discretized and vectorized to digital vectors. Finally, digital clustering and hierarchical clustering with Pearson's correlation using the group average method were applied to the genes.

Results

Benchmark testing

We evaluated the accuracy of digital clustering and the conventional hierarchical clustering method with Pearson 's correlation using the group average method for the simulated clusters. For all pairs of genes that have identical labels, we assessed whether they belong to the same cluster after clustering. Genes that had identical labels and belonged to the same cluster were true positives. Genes that had different labels and belonged to different clusters were true negatives. Genes that had the same label and belonged to different clusters were considered false-negatives. Genes that had different labels and belonged to the same cluster were considered false-positives. Thus, we converted the clustering problem to a binary classification problem. The receiver operating characteristic (ROC) curve was created by plotting the true positive rate (TPR; true positive / condition positive) against the false-positive rate (FPR; false positive / condition negative) at the merge step of clustering for each iteration. The area under the ROC curve (AUC) was used as an indicator in the evaluation.

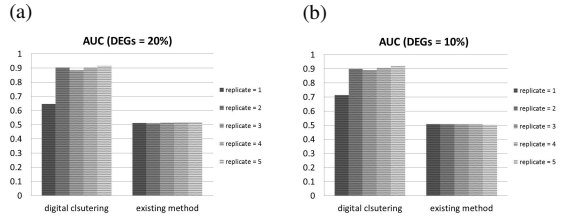

Fig. 1 Evaluation of clustering.

Figure 1 summarizes the evaluation of the clustering methods. Three simulations were performed for each combination of parameters, and each vertical bar in the figure represents the average AUC for the three simulations. The AUC for the existing method was approximately 0.5 and did not depend on the number of replicates or the proportion of DEGs. Our method not only showed significantly higher accuracy than that of the existing method for all parameter values, but the AUCs also improved when there were two or more replicates.

Conclusion

We developed a novel clustering algorithm, digital clustering, based on statistical test results. It demonstrated high classification accuracy when applied to multi-conditional transcriptome data. Digital clustering can be applied to any multi-conditional transcriptome data, and to the results of any DEG detection tool given an appropriate input format. Accordingly, it has a wide range of applications for transcriptome data analysis.

References

[1] Metzker, M. L.: Sequencing technologies - the next generation., *Nat Rev Genet*, Vol. 11, No. 1, pp. 31–46 (online), DOI: 10.1038/nrg2626 (2010).

[2] Wang, Z., Gerstein, M. and Snyder, M.: RNA-Seq: a revolutionary tool for transcriptomics., *Nat Rev Genet*, Vol. 10, No. 1, pp. 57–63 (online), DOI: 10.1038/nrg2484 (2009).

[3] D'haeseleer, P.: How does gene expression clustering work?, *Nat Biotechnol*, Vol. 23, No. 12, pp. 1499–1501 (online), DOI: 10.1038/nbt1205-1499 (2005).

[4] Elowitz, M. B., Levine, A. J., Siggia, E. D. and Swain, P. S.: Stochastic gene expression in a single cell., *Science*, Vol. 297, No. 5584, pp. 1183–1186 (online), DOI: 10.1126/science.1070919 (2002).

[5] Pickrell, J. K., Marioni, J. C., Pai, A. A., Degner, J. F., Engelhardt, B. E., Nkadori, E., Veyrieras, J.-B., Stephens, M., Gilad, Y. and Pritchard, J. K.: Understanding mechanisms underlying human gene expression variation with RNA sequencing., *Nature*, Vol. 464, No. 7289, pp. 768–772 (online), DOI: 10.1038/nature08872 (2010).

[6] Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L. and Pachter, L.: Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks., *Nat Protoc*, Vol. 7, No. 3, pp. 562–578 (online), DOI: 10.1038/nprot.2012.016 (2012).