ROKU: An Improved Method for the Detection of Tissue-Specific Expression Patterns

Koji Kadota  
Jiazheng Ye  
Yuji Nakai  
Kadota@iu.a.u-tokyo.ac.jp  
ye@bi.a.u-tokyo.ac.jp  
yunakai@iu.a.u-tokyo.ac.jp  
Tohru Terada  
Kentaro Shimizu  
terada@iu.a.u-tokyo.ac.jp  
shimizu@bi.a.u-tokyo.ac.jp  

Graduate School of Agricultural and Life Sciences, University of Tokyo, 1-1-1 Yayoi, Bunkyo-ku, Tokyo 113-8657, Japan

Keywords: microarray, differential expression, entropy, AIC

1 Introduction

A major challenge of microarray analysis is to detect genes whose expression in a single or small number of tissues is significantly different than in other tissues. The tissue-specificity provides an essential reference to identify tissue-specific drug targets.

Several methods were used for the purpose: For example, Shannon information theoretic entropy was used for ranking genes according to their tissue-specificity [1]; Akaike’s information criterion (AIC) was used for the detection of expression outliers [2]. Both the entropy- and AIC-based methods entail pros and cons. The entropy does not explain to which tissue a gene is tissue-specific, only measuring the degree of overall tissue specificity of the gene. The AIC-based method is the contrary.

This complementary relationship between the two methods led us to develop a combined approach, ROKU. ROKU analyzes any type of tissue-specific genes (up-, down-, and mixed-type) in two steps. First, it ranks genes according to overall tissue-specificity using entropy [1], and second, for each gene, it identifies specific tissues whose observations are regarded as outliers using AIC [2].

2 Method

2.1 Gene Ranking by Shannon Entropy

Consider one gene’s expression vector \( x = (x_1, x_2, \ldots, x_N) \) for \( N \) tissues and an observation \( x_t \) for tissue \( t \). The entropy of the gene is calculated as 
\[
H = -\sum_{i=1}^{N} p_i \log_2(p_i),
\]
where \( p_i \) is the relative expression of \( x_i \) for tissue \( t \) defined as 
\[
p_i = x_i / \sum_{i=1}^{N} x_i.
\]

\( H \) ranges from zero to \( \log_2(N) \), with the value 0 for genes expressed in a single tissue (Fig. 1a) and \( \log_2(N) \) for genes expressed uniformly in all the interrogated tissues (Fig. 1b). We therefore rely on the low entropy score for the identification of tissue-specific genes.

To equally identify down- and mixed-types of tissue-specific genes as well as up-type genes, we processed the original vector \( x \). The processed observation \( x_t' \) for tissue \( t \) is defined as 
\[
x_t' = |x_t - T_{bw}|,
\]
where \( T_{bw} \) is the one-step Tukey biweight.

2.2 Detecting Specific Tissues as Outliers

As mentioned above, the entropy does not indicate which tissues are specific, but is a measure of the overall tissue specificity of a gene. We imagine observations in specific tissues to be easily visualized as outliers on the over- and/or under-expressed side if any exist. We used an AIC-based outlier detection method [2] to detect tissues with specific expression patterns.
3 Results and Discussion

We here divided tissue-specific genes into two levels, a narrow sense and a broad sense. Genes over-expressed in a small number of tissues but unexpressed or slightly expressed in others, such as those shown in Figs. 1a and c, are defined as tissue-specific genes in a narrow sense, while genes over- and/or under-expressed in a small number of tissues compared to other tissues are defined as tissue-specific in a broad sense (or tissue-selective).

Figure 1 are synthetic expression observations for \( N = 10 \) tissues. The entropy \( H \) for each gene vector \( x \) is given by the number in black above the figures. Clearly, direct calculation of the entropy for raw gene vector \( x \) works well only for detecting tissue-specific genes in a narrow sense (Figs. 1a and c) but not for those in a broad sense (Figs. 1d-f). The \( H \) scores \((3.22, 3.29, 3.23)\) for Figs. 1d-f, respectively) of tissue-specific genes in a broad sense are close to the maximum value \((\log_{10} 10 = 3.32)\) and cannot identify those genes as tissue-specific.

To detect tissue-specific genes in a broad sense, ROKU processes a given gene vector \( x \) and makes a new vector \( x' \). The scatter plots of processed vectors are shown in gray in Fig. 1. The entropy scores \( H(x') \) for the processed vectors to obvious tissue-specific genes in a broad sense \((1.74, 1.48, 1.64)\) of Figs. 1d-f, respectively) are considerably lower than those for \( x \). This is because the relative expression levels for specific tissues (highlighted tissues detected by AIC [2]) become high after data processing. These results demonstrate the adequacy for our strategy for detecting tissue-specific genes in a broad sense.

References

