Detection of Genes with Tissue-Specific Patterns Using Akaike's Information Criterion

Koji Kadota koji-kadota@aist.go.jp
Katsutoshi Takahashi takahashi-k@aist.go.jp

Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology, 2-43 Aomi Koto-ku, Tokyo 135-0064, Japan

Keywords: outlier detection, microarray, expression analysis, AIC

1 Introduction

One of the important challenges of microarray analysis is the identification of tissue-specific genes whose expression profile is considerably different in particular tissue(s) than in others. Those characteristics facilitate the identification of a large number of possible markers. In general, the problem of identifying tissue-specific expression patterns in multisource data can be viewed as an outlier identification problem.

Akaike's Information Criterion (AIC), introduced almost 30 years ago by H. Akaike, is an information criterion for the identification of an optimal model from a class of competing models [1]. Kitagawa [4] subsequently used AIC to detect outliers and Ueda [5] more recently simplified the procedure. The most significant advantages of those methods are (i) it is possible to reach a relatively objective decision because the procedure does not require the selection of a significance level, (ii) various situations (e.g. single outlier, multiple lowest or highest outliers, two-sided- and grouped cases) can be treated equally.

We report the application of a simplified method for the identification of markedly contracting genes from microarray data [2]. The validity of this novel approach is demonstrated by the distribution of the data detected as outliers and by the comparison with the other method.

2 Methods

According to Ueda [5], a statistic U to identify outliers is defined as

$$U = \frac{1}{2}AIC = n\log\sigma + \sqrt{2} \times s \times \frac{\log n!}{n}$$

where $(n+s)$ denotes the total number of observations, s denotes the number of outlier candidates, and s denotes the standard deviation of scores assigned to n samples excluding outlier candidates.

The statistic U has a clear interpretation in outlier detection. A low value for the first term in the equation does, while a high value does not, indicate that the combination of s outlying observations is likely to be bona fide. The second term indicates increased unreliability due to an increased number of parameters (in this case, s). The best approximating combination is one that achieves the lowest value for U and is termed the Minimum AIC Estimate (MAICE). The procedure aimed at obtaining the MAICE of the models is called the minimum AIC procedure [4].

Consider, for example, a series of observations of '2.1', '2.5', '3.2', '3.4', '8.6', and '9.7'. We expect the last two observations to be identified as outliers. The MAICE obtained from the minimum AIC procedure is indeed same as the intuition (see Table 1).

Table 1: The minimum AIC procedure. The MAICE corresponds to Model3.

Combination	2.1	2.5	3.2	3.4	8.6	9.7	*U(AIC/2)*
Model1							-0.5
Model2						outlier	-0.4
Model3					outlier	outlier	-5.2
Model4				outlier	outlier	outlier	-3.5
Model5	outlier						0.9
Model6	outlier					outlier	1.0
Model7	outlier				outlier	outlier	-4.0
Model8	outlier	outlier					1.7
Model9	outlier	outlier				outlier	1.6
Model10	outlier	outlier	outlier				1.9

3 Results and Discussion

We applied the minimum AIC procedure to detect outliers as tissue-specific patterns in publicly available gene expression data consisting of a large variety of tissues. As a result, we found that the current procedure could extract specific expression patterns from arbitrarily selected tissues, lung-specific patterns, for example. We also found some advantageous characteristics of the method compared to the other method such as pattern-matching. The former seemed to be able to detect brain-specific genes (highly expressed only in "brain" and "cerebellum") while the latter also included some extra observations especially in "cortex" and "eyeball". Accordingly, we conclude that the minimum AIC procedure is specifically applicable to the extraction of specific expression patterns from arbitrarily selected tissues under the condition of co-existing similar tissues.

The advantages of the method we proposed here are (i) the acquired answer is objective and (ii) various situations (e.g. single outlier, multiple lowest or highest outliers, two-sided- and grouped cases) can be treated equally. As these characteristics mirror those of the method currently in wide use, our method appears to be readily applicable to various expression data [3].

References

[1] Akaike, H., Information theory and an extension of the maximum likelihood principle, *2nd International Symposium on Information Theory*, 267–281, 1973.

[2] Kadota, K., Nishimura, S.I., Bono, H., Nakamura, S., Hayashizaki, Y., Okazaki, Y., and Takahashi, K., Detection of genes with tissue-specific expression patterns using Akaike's Information Criterion procedure, *Physiol. Genomics*, 12(3):251–259, 2003.

[3] Kadota, K., Tominaga, D., Akiyama, Y., and Takahashi, K., Detecting outlying samples in microarray data: A critical assessment of the effect of outliers on sample classification, *Chem-Bio Informatics J.*, 3(1):30–45, 2003.

[4] Kitagawa, G., On the use of AIC for the detection of outliers, *Technometrics*, 21:193–199, 1979.

[5] Ueda, T., Simple method for the detection of outliers, *Japanese J. Appl. Stat.*, 25:17–26, 1996.