Estimation of the Optimal Stage Division on Gene Expression Time Series

Daisuke Tominaga
Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology
2-4-7 Aomi, Koto, Tokyo 135-0064, Japan
tominaga@cbrc.jp

Background

In biological phenomena at the cellular level, such as cancer or embryo development, changes of state, phase or condition of cells are observed as stage progression, and gene expression levels of these cells may change in association with these stages. A statistical distribution model can represent the distribution of a gene's expression over time in a particular stage. Given quantitative time series data, the optimal model for differentiation into stages is defined by the total likelihood of distribution models for all stages of a series[1]. However, determination of the optimal number and times of distinctions between different stages is difficult.

Method

We introduce the Akaike Information Criterion (AIC) to tackle this problem. AIC is a scalar value that is minimal for the optimal mathematical model that fits best to the given data with the least number of free parameters of the model on the assumption that the noise of the data has a normal distribution. In the case of gene expression levels determined by DNA microarray, the distribution of observed values is log-normal. Therefore, the IC can be applied to log values of microarray data. An exhaustive search for all potential models for division into stages can find the best (AIC minimal) model (Figure 1). This optimal model for division into stages is found for a time series of each gene, and we can then count genes whose states in terms of expression level change at every interval between sampling points.

Results

We tested the reliability of this approach by applying it to simulated gene expression time series and experimentally observed data of early developmental stages of nematode[2]. The results from simulation data showed how our algorithm correctly identified times at which one stage changed to another. We generated 500 time series that consisted of ten sampling points and two stages. Data values in each stage were normally distributed random numbers. Variances were the same in both stages while means differed. The accuracy of our algorithm was shown to vary with the ratio of the mean to the variance (Figure 2A).

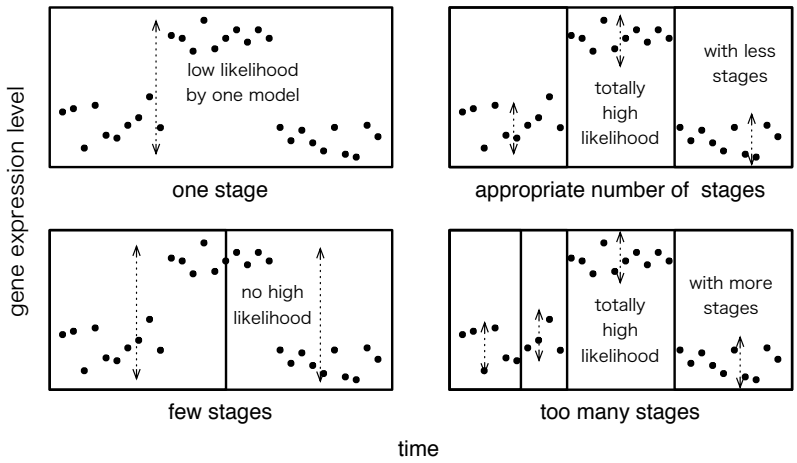

Figure 1. Optimal stage division at the viewpoint of the information criterion.

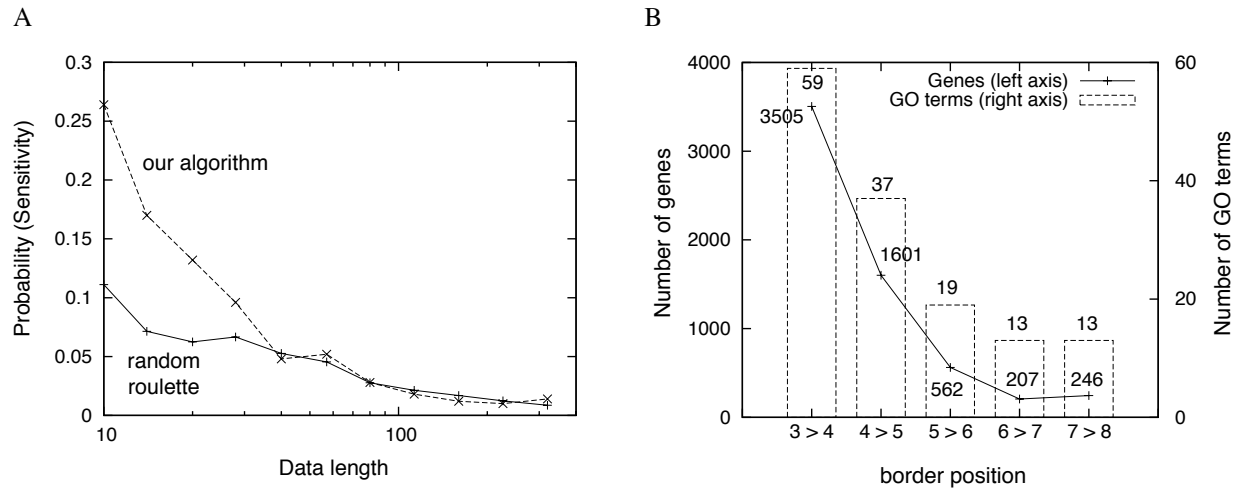

Figure 2. (A) Comparison of our algorithm for optimal stage division to random division, and (B) numbers of up-regulated genes at the estimated stage borders (solid line) and numbers of significant GO terms of these genes (box).

Nematode data consisted of ten sampling points for each gene. We chose 1615 genes excluding duplicated genes. Our algorithm identified the timing and a number of stage transitions of expression levels for each gene. We counted the number of genes that changed stage for each interval between sampling time points, and calculated p-values of Gene Ontology (GO) terms that were used to annotate to these genes. GO term analyses showed good agreements with the literature[3] (Figure 2B).

(1) Yi, M., *et. al.*, Novel Staging System for Predicting Disease-Specific Survival in Patients With Breast Cancer Treated With Surgery as the First Intervention: Time to Modify the Current American Joint Committee on Cancer Staging System, *Journal of Clinical Oncology*, **35**:4654-4661 (2011).
(2) Baugh, L.R., *et. al.* Composition and dynamics of the *Caenorhabditis elegans* early embryonic transcriptome, *Development*, **130**:889-900 (2003).
(3) Baugh, L.R., *et. al.* The homeodomain protein PAL-1 specifies a lineage-specific regulatory network in the *C. elegans* embryo, *Development*, **132**:1843-1854 (2005).