Mathematical Biosciences 254 (2014) 58–63

Contents lists available at ScienceDirect

Mathematical Biosciences

journal homepage: www.elsevier.com/locate/mbs

Statistical stage transition detection method for small sample gene expression time series data

Daisuke Tominaga *

Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology (AIST), 2-4-7 Aomi, Koto, Tokyo 135-0064, Japan

ARTICLE INFO

Article history:
Received 17 September 2013
Received in revised form 3 June 2014
Accepted 4 June 2014
Available online 21 June 2014

Keywords:
Information criterion
Caenorhabditis elegans
Development
Gene ontology

ABSTRACT

In terms of their internal (genetic) and external (phenotypic) states, living cells are always changing at varying rates. Periods of stable or low rate of change are often called States, Stages, or Phases, whereas high-rate periods are called Transitions or Transients. While states and transitions are observed phenotypically, such as cell differentiation, cancer progression, for example, are related with gene expression levels. On the other hand, stages of gene expression are definable based on changes of expression levels. Analyzing relations between state changes of phenotypes and stage transitions of gene expression levels is a general approach to elucidate mechanisms of life phenomena.

Herein, we propose an algorithm to detect stage transitions in a time series of expression levels of a gene by defining statistically optimal division points. The algorithm shows detecting ability for simulated datasets. An annotation based analysis on detecting results for a dataset of initial development of *Caenorhabditis elegans* agrees with that are presented in the literature.

1. Introduction

Internal (genetic) and external (phenotypic) states in living cells are changing continuously. Rates of change are also changing. Generally states of living cells are classifiable into two kinds: stable states and transitions. Stable states are also designated as 'stages' or 'phases', and are found in many biological phenomena such as individual development [1], cancer progression [2], fermentation [3,4], and actions of individuals [5]. Reports on these studies show that stages defined by phenotypic changes are related with changes in gene expression levels. Therefore, defining stages based on gene expression levels are presumably consistent with phenotypic stages.

Applications of statistical distribution models to gene expression time series for stage analysis are described in many reports. For example, determination of starting timing of gene expression [6], identification of gene regulatory networks based on gene expression time series divided by given stage definitions [7], and comparing plural division patterns of gene expression time series of cancer cells [8] based on the Information Criterion [9] are recent works. These methods are based on pre-defined stage information on time series data. Stage information is based on observed phenotypic state changes or general analyses of transitions in phenomena.

We here are interested in how to find stages statistically in a time series of expression levels of a gene without any other information. We expected that these statistically detected stages of a gene agree with phenotypic states. Reports introduced above suggest that existence of certain relationships between phenotypic states and gene expression stage transitions. We think that gene expression stages which are found without phenotypic information possibly agree with phenotypic changes.

Generally, statistical tests necessitate the use of a large sample, *e.g.*, tests for normality of a distribution require 20–30 samples or more for reliable conclusions [10]. Contrary to this, publicly available gene expression time series datasets often consist of a few samples. Other statistical methods for stage transitions, 'Change point detection' methods, present similar disadvantages. Methods of this kind judge whether a new sample has come from the same population of previously obtained samples. Therefore a sufficient number of samples as previous samples are needed [11,12].

We propose a new algorithm to determine statistically optimal division points on quantitative gene expression time series data into stages. The algorithm is based on brute force search for all possible division patterns that consist of the number and the positions of division points. For time series observations using DNA microarrays, the algorithm can determine the stage transition points for all each genes. Therefore, for each interval between sapling time points, the number of genes that has a stage transition point on the interval can be counted. State changes in a whole cell can be regarded as a summary of all stage transitions of each gene.

* Tel.: +81 3 3599 8080; fax: +81 3 3599 8081.
E-mail address: tominaga@cbrc.jp

http://dx.doi.org/10.1016/j.mbs.2014.06.005
0025-5564/

We measured the detection accuracy of our method through its application to simulated datasets. Thus we applied our algorithm to the gene expression dataset of initial development of *Caenorhabditis elegans* where phenotypic states are defined clearly as phases of the egg cleavage. Annotation-based analyses of statistically optimal stages on gene expression levels show good agreement with literal knowledge of phenotypic stages.

2. Material and methods

2.1. Assumptions on quantitative gene expression data

Our algorithm is based on several assumptions on quantitative gene expression time series data, and we suppose that these data are obtained by DNA microarray observations. The first is that a time series of expression levels of a gene consists of one or more stages. The second is that values of gene expression levels can be modeled statistically by the normal distribution. This assumption can be controversial because no model is established today for distributions of gene expression levels and the distribution do not seem to normal in some cases [13,14]. Here we introduce the assumption to demonstrate how the algorithm work. And a report about log-normality of the distribution [15] supports this assumption. The third assumption is that stages transit within an interval between sampling time points.

2.2. Exhaustive search for division patterns of time series data

The algorithm searches the statistically optimal stages division positions in quantitative time series data using an exhaustive pattern search for all possible numbers and positions of stage divisions. Input to the algorithm is a sequence of real numbers as a time series of expression levels of a gene, and output is a 'division pattern,' a set of optimal dividing positions on the series into stages, namely the numbers and the positions of stage borders. The information criterion value is calculated for each of all mathematically possible division patterns based on the number of division positions and the total likelihood values of statistical distribution models those are fitted for each stage. Better fit with fewer stages is evaluated better by the information criterion (Fig. 1). The optimal division pattern is of the minimum information criterion value. Stages in the optimal pattern are regarded as 'detected'.

2.3. Calculation of AIC

The information criterion (we use here Akaike's Information Criterion, AIC [9]) is calculated from the sum of the degree of freedom and log-likelihood of the model. The model is the division pattern. The degree of freedom is the number of division points in the division pattern (the number of stages minus one). The likelihood value is calculated as the total likelihood of all each statistical distribution model for each stage in the division pattern. We use the normal distribution model here for each stage. The total likelihood of the model is a product of all likelihood values of each distribution model for each stage. Therefore the AIC value for the model is calculated as

$$\frac{1}{2}\mathrm{AIC} = s - 1 + \log\left(\prod_{i=1}^{s} D_i\right), \tag{1}$$

where s is the number of stages of the model and D_i is the likelihood of the statistical distribution model for the i-th stage in the model, which is calculated as follows:

$$D_i = \prod_{k=1}^{n_i}\left\{\frac{1}{\sqrt{2\pi}\bar{\sigma}_i}\exp\left(-\frac{(x_{ik}-\bar{\mu}_i)^2}{2\bar{\sigma}_i^2}\right)\right\}, \tag{2}$$

where n_i is the number of samples in the i-th stage of the model, $\bar{\sigma}_i$ is the standard deviation of samples (expression levels of a gene) of the i-th stage, x_{ik} is the value of the k-th sample in the i-th stage, and $\bar{\mu}_i$ is the sample mean value of the i-th stage. The AIC value is calculated for all possible models on the given time series data under the condition; the minimal length of a stage L_{min} that is defined by users.

s in Eq. (1) varies for division patterns. Let L is the length of given time series data ($L = \sum_{i=1}^{s} n_i$), the maximum value of s, s_{max}, is the maximum integer value that is less than L/L_{min}. The possible length of a stage is in a range of L_{min} to L.

Our algorithm searches the optimal s and the optimal positions of dividing points on the given sequence of real numbers.

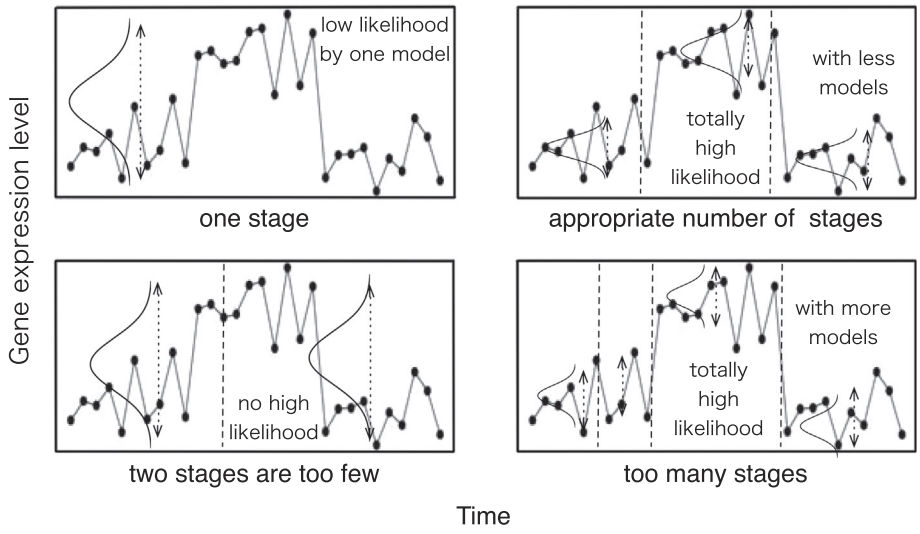

Fig. 1. Optimal division of time series of quantitative data into stages. Sample distributions in a time series data are modeled by statistical distribution models (the normal distribution model is applied in this paper). Likelihood values of fitted statistical models are the highest when the series of samples are divided appropriately into groups (stages). The best division is of higher fitness values with the fewer number of divisions. The total likelihood of a division pattern is the product of those of all fitted statistical distribution models. The information criterion values that represents the goodness of the division pattern is calculated from the total likelihood and the degrees of freedom of a division pattern that is equal to the number of division points.