An Effective Approach for Mining Time-Series Gene Expression Profile

Vincent S. M. Tseng, Yen-Lo Chen

Department of Computer Science and Information Engineering, National Cheng Kung University, *Tainan, Taiwan, R.O.C.*

(Email: tsengsm@mail.ncku.edu.tw)

Abstract. Time-series data analysis is an important problem in data mining fields due to the wide applications. Although some time-series analysis methods have been developed in recent years, they can not effectively resolve the fundamental problems in time-series gene expression mining in terms of scale transformation, offset transformation, time delay and noises. In this paper, we propose an effective approach for mining time-series data and apply it on time-series gene expression profile analysis. The proposed method utilizes dynamic programming technique and correlation coefficient measure to find the best alignment between the time-series expressions under the allowed number of noises. Through experimental evaluation, our method was shown to effectively resolve the four problems described above simultaneously. Hence, it can find the correct similarity and imply biological relationships between gene expressions.

1 Introduction

Time-series data analysis is an important problem in data mining with wide applications like stock market analysis and biomedical data analysis. One important and emerging field in recent years is mining of time-series gene expression data. In general, gene expression mining aims at analysis and interpretation of gene expressions so as to understand the real functions of genes and thus uncover the causes of various diseases [3, 8, 9, 20, 21, 29, 30, 31]. Since the gene expression data is in large scale, there is a great need to develop effective analytical methods for analyzing and exploiting the information contained in gene expression data. A number of

relevant studies have shown that cluster analysis is of significant value for the exploration of gene expression data [3, 4, 11-13, 20, 21].

Although a number of clustering techniques have been proposed in recent years [9, 10, 13, 20, 21-23], they were mostly used for analyzing multi-conditions microarray data where the gene expression value in each experimental condition is captured only at a time point. In fact, biological processes have the property that multiple instances of a single process may unfold at different and possibly non-uniform rates in different organisms or conditions. Therefore, it is important to study the biological processes that develop over time by collecting RNA expression data at selected time points and analyzing them to identify distinct cycles or waves of expression [3, 4, 13, 25, 30]. In spite that some general time-series analysis methods were developed in the past decades [2, 4, 14-18], they were not suited for analyzing gene expressions since the biological properties were not considered.

In the time-series gene expression data, the expression of each gene can be viewed as a curve under a sequence of time points. The main research issue in clustering time-series gene expression data is to find the similarity between the time-series profiles of genes correctly. The following fundamental problems exist in finding the similarity between time-series gene expressions:

Scaled and offset transformations: For two given time-series gene expressions, there may exist the relations of scaled transformation (as shown in Figure 1a) or offset transformation (as shown in Figure 1b). In many biological applications based on gene expression analysis, genes whose time-series expressions are of scaled or offset transformation should be given high similarity since they may have highly-related biological functions. Obviously, frequently used measures in clustering like “distance” can not work well for these transformation problems.

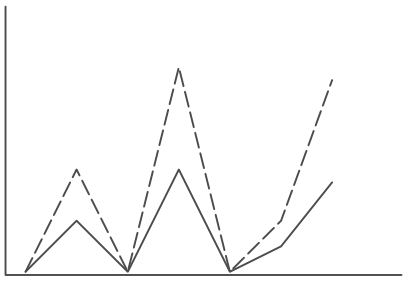

Fig. 1.a. Scaled transformation