Available online at www.sciencedirect.com
SCIENCE DIRECT®
Fuzzy Sets and Systems 152 (2005) 49–66
FUZZY sets and systems
www.elsevier.com/locate/fss

Clustering of unevenly sampled gene expression time-series data

C.S. Möller-Levet[a], F. Klawonn[b], K.-H. Cho[c,d,*], H. Yin[a], O. Wolkenhauer[e,1]

[a]*Department of Electrical Engineering and Electronics, University of Manchester, Institute of Science and Technology, Manchester M60 1QD, UK*
[b]*Department of Computer Science, University of Applied Sciences, D-38302 Wolfenbüttel, Germany*
[c]*College of Medicine, Seoul National University, Chongno-gu, Seoul, 110-799, Republic of Korea*
[d]*Korea Bio-MAX Center, Seoul National University, Gwanak-gu, Seoul, 151-818, Republic of Korea*
[e]*Department of Computer Science, Systems Biology & Bioinformatics Group, University of Rostock, Albert-Einstein Str. 21, 18059 Rostock, Germany*

Available online 15 December 2004

Abstract

Time course measurements are becoming a common type of experiment in the use of microarrays. The temporal order of the data and the varying length of sampling intervals are important and should be considered in clustering time-series. However, the shortness of gene expression time-series data limits the use of conventional statistical models and techniques for time-series analysis. To address this problem, this paper proposes the fuzzy short time-series (FSTS) clustering algorithm, which clusters profiles based on the similarity of their relative change of expression level and the corresponding temporal information. One of the major advantages of fuzzy clustering is that genes can belong to more than one group, revealing distinctive features of each gene's function and regulation. Several examples are provided to illustrate the performance of the proposed algorithm. In addition, we present the validation of the algorithm by clustering the genes which define the model profiles in Chu et al. (Science, 282 (1998) 699). The fuzzy *c*-means, *k*-means, average linkage hierarchical algorithm and random clustering are compared to the proposed FSTS algorithm. The performance is evaluated with a well-established cluster validity measure proving that the FSTS algorithm has a better performance than the compared algorithms in clustering similar rates of change of expression in successive unevenly distributed time points. Moreover, the FSTS algorithm was able to cluster in a biologically meaningful way the genes defining the model profiles.

Keywords: Fuzzy clustering; Unevenly sampled; Short time series; Gene expression

* Corresponding author. College of Medicine, Seoul National University, Chongno-gu, Seoul, 110-799, Republic of Korea. Tel.: +82 2 887 2650; fax: +82 2 887 2692.
E-mail address: ckh-sb@snu.ac.kr (K.-H. Cho).
[1] Also for correspondence.

0165-0114/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.fss.2004.10.014

1. Introduction

Microarrays revolutionise the traditional way of one gene per experiment in molecular biology [4]. With microarray experiments it is possible to measure simultaneously the activity levels of thousands of genes. An appropriate clustering of gene expression data can lead to meaningful classification of diseases, identification of co-expressed functionally related genes, logical descriptions of gene regulation, etc.

Time course measurements are becoming a common type of experiment in the use of microarrays. The particularity of time-series, which has to be considered in the clustering analysis, is the temporal information: the measurements ordered in time and sampled at specific intervals. An appropriate similarity measure for gene expression time-series should be able to identify similar shapes which are formed by the relative change of expressions in temporal information.

This paper is organised as follows: the effects of the temporal information in the comparison of shapes are discussed first, followed by the related work. The next section defines the short time-series (STS) distance, develops the fuzzy short time-series (FSTS) algorithm using the standard fuzzy *c*-means algorithm (FCM) as a template, and provides simple examples to demonstrate its performance. Then, the validation of the algorithm is presented by clustering the genes which define the model profiles in [5]. The fuzzy *c*-means, *k*-means, average linkage hierarchical algorithm and random clustering are used for comparison. A well-established validity measure relevant to gene expression clustering is applied to evaluate the quality of the clusters. In addition, the results are discussed using the external biological criteria. Then, the scopes and limitations of the FSTS algorithm are discussed. Finally, conclusions are presented in the final section summarising the presented research.

2. Temporal information and clustering

To visualise the effects of the temporal information in the comparison of shapes consider the following example. The microarray analysis of *Saccharomyces cerevisae* by Chu et al. [5] shows that PYC1 and SIP4 are two of the 52 genes that were induced rapidly and transiently after transfer to sporulation medium. PYC1 is involved in the gluconeogenesis pathway as a pyruvate carboxylase and SIP4 is a transcription factor, which interacts with the SNF1 protein kinase. These genes were part of the handpicked genes selected for the "metabolic'' model profile used in [5]. Consider a synthetic gene (GENEX), whose standardised [1] expression values are identical to those of SIP4 except for the first time point which has a higher expression. Fig. 1(a) illustrates the resulting profile along with the standardised values of PYC1 and SIP4. When comparing the similarity of SIP4 and GENEX to PYC1, it can be observed that PYC1 and SIP4 have a more similar induction period after transfer to sporulation medium. That is, the relative change of expression from the first to the second measurement of PYC1 is more similar in SIP4 than in GENEX, while all the other changes are equal in SIP4 and in GENEX. However, when using the Euclidean distance to assess the similarity, PYC1 is more similar to GENEX ($d_E = 1.45$) than to SIP4 ($d_E = 1.57$). The Euclidean distance is invariant with respect to the order of the observations, therefore, the direction of change of expression (i.e. up–down) is not considered. The next element to consider is the length of sampling intervals. By including this not only the direction of the change is considered but also the rate of change. Biological processes are sampled at shorter intervals of time when intense

[1] Standardised values (zero mean and standard deviation of one) are utilised to eliminate shifting and scaling factors.