Infectious Disease Modelling 1 (2016) 28–39

Contents lists available at ScienceDirect

Infectious Disease Modelling

journal homepage: www.keaipublishing.com/idm

Correlation-based iterative clustering methods for time course data: The identification of temporal gene response modules for influenza infection in humans

Michelle Carey [a, b], Shuang Wu [a, c], Guojun Gan [d], Hulin Wu [a, e, *]

[a] Department of Biostatistics and Computational Biology, Crittenden Blvd, Rochester, NY 14642, USA
[b] Department of Mathematics and Statistics, McGill University, 805 Sherbrooke Street West, Montreal, Canada
[c] Biogen, 250 Binney Street, Cambridge, MA, USA
[d] Department of Mathematics, University of Connecticut, 196 Auditorium Road U-3009, Storrs, USA
[e] Department of Biostatistics, University of Texas Health Science Center School of Public Health at Houston, 1200 Pressler Street, Houston, USA

ARTICLE INFO

Article history:
Received 9 May 2016
Accepted 8 July 2016
Available online 2 September 2016

Keywords:
Clustering
Inhomogeneous clusters
Power law

ABSTRACT

Many pragmatic clustering methods have been developed to group data vectors or objects into clusters so that the objects in one cluster are very similar and objects in different clusters are distinct based on some similarity measure. The availability of time course data has motivated researchers to develop methods, such as mixture and mixed-effects modelling approaches, that incorporate the temporal information contained in the shape of the trajectory of the data. However, there is still a need for the development of time-course clustering methods that can adequately deal with inhomogeneous clusters (some clusters are quite large and others are quite small). Here we propose two such methods, hierarchical clustering (IHC) and iterative pairwise-correlation clustering (IPC). We evaluate and compare the proposed methods to the Markov Cluster Algorithm (MCL) and the generalised mixed-effects model (GMM) using simulation studies and an application to a time course gene expression data set from a study containing human subjects who were challenged by a live influenza virus. We identify four types of temporal gene response modules to influenza infection in humans, i.e., single-gene modules (SGM), small-size modules (SSM), medium-size modules (MSM) and large-size modules (LSM). The LSM contain genes that perform various fundamental biological functions that are consistent across subjects. The SSM and SGM contain genes that perform either different or similar biological functions that have complex temporal responses to the virus and are unique to each subject. We show that the temporal response of the genes in the LSM have either simple patterns with a single peak or trough a consequence of the transient stimuli sustained or state-transitioning patterns pertaining to developmental cues and that these modules can differentiate the severity of disease outcomes. Additionally, the size of gene response modules follows a power-law distribution with a consistent exponent across all subjects, which reveals the presence of universality in the underlying biological principles that generated these modules.

* Corresponding author. Department of Biostatistics, University of Texas Health Science Center School of Public Health at Houston, 1200 Pressler Street, Houston, USA.
E-mail address: Hulin.Wu@uth.tmc.edu (H. Wu).
Peer review under responsibility of KeAi Communications Co., Ltd.

http://dx.doi.org/10.1016/j.idm.2016.07.001

1. Introduction

Understanding a host temporal response to a disease is imperative to the development of predictive and preventive medicine. Many diseases have critical transition points that are linked to the severity of disease outcomes (Li, Jin, Lei, Pan, & Zou, 2015; Rietkerk, Dekker, de Ruiter, & van de Koppel, 2004; Yu, Li, & Chen, 2014). Identifying these points and deciphering which genes are biomarkers for predicting these transitions is a challenging biological problem. Time course gene expression data provides a description of the dynamic features of the gene-level response to a disease and/or external stimulation. Genes with similar temporal response patterns can be grouped together to form temporal gene response modules.

Many clustering methods have been developed to identify temporal gene response modules for time course data. However, most of these methods do not incorporate the fact that there are many co-expressed or redundant genes that follow similar temporal patterns, but at the same time, there are genes with very few or even no co-expressed or redundant genes, and thus exhibit unique temporal response patterns. Consequently, the temporal gene response modules or clusters can be inhomogeneous, i.e., some clusters are very large and contain many genes while others are small or even only contain a single gene. For example the majority of the standard methods for clustering vectors, including high-dimensional vectors, such as centre-based clustering methods e.g., k-means (Hartigan & Wong, 1979), hierarchical clustering (Eisen, Spellman, Brown, & Botstein, 1998), graph-based or grid-based algorithms, model-based approaches and self-organizing maps (SOM) (Kohonen, 1995), see (Gan, Ma, & Wu, 2007) for a detailed review, are not flexible enough to deal with inhomogeneous clusters.

Density-based spatial clustering of applications with noise (DBSCAN) (Ester, Kriegel, Sander, & Xu, 1996) and model-based hierarchical clustering (Fraley and Raftery, 2002) have been proposed to identify clusters of different sizes, shapes and densities, although these methods cannot simultaneously identify both very large and very small (even single-element) clusters. As a result, mixture modelling approaches and iterative clustering algorithms have been introduced to identify the rare events or small cell populations from flow cytometry data (Chan & Vasconcelos, 2008; Cron et al., 2013; Naim, Datta, RebhahnCavenaugh, Mosmann, & Sharma, 2014). However, the computational implementation of these methods is expensive and the algorithms may not easily identify a consistent cluster assignment for large gene expression data sets with complex temporal structures.

Autoregressive time series models (Ramoni, Sebastiani, & Cohen, 2002a, 2002b), hidden Markov models (Schliep, Schönhuth, & Steinhoff, 2003) and generalised mixed-effects models (GMM) (Bar-Joseph, 2004; Lu, Liang, Li, & Wu, 2011; Ma, Castillo-Davis, Zhong, & Liu, 2006) have been proposed to cluster time course gene expression data. Typically, these procedures utilise a Bayesian clustering framework equipped with either a Markov chain Monte Carlo or an EM algorithm, which tend to be computationally intensive with a very slow convergence. Additionally, these algorithms require robust initial estimates of the model parameters which are often difficult to attain.

Graph theory approaches, namely, the Markov Cluster Algorithm (MCL) (van Dongen, 2000a,b), the MCODE algorithm (Bader & Hogue, 2003), restricted neighbourhood search clustering (RNSC) (Douglas King, 2004) and super paramagnetic clustering (SPC) (King, Przulj, & Jurisica, 2004), which exploit the local structure in networks, have experienced an increase in popularity in the literature. This is mainly due to the simplicity of the algorithms, the automatic selection of the number of clusters and their capacity to identify inhomogeneous clusters. (Brohee and van Helden, 2006) show that in general the MCL outperforms the MCODE algorithm, RNSC and SPC. Furthermore, the MCL has been shown to be an apt method for identifying novel aspects of biological functions for gene expression data (Freeman et al., 2007).

In this article, we propose to use correlation-based iterative clustering methods to effectively identify inhomogeneous clusters from time course gene expression data. We expect that our methods will provide a more reliable approach for the identification of temporal gene response modules in comparison to the graph theory and mixture model approaches. This will be demonstrated by applying the proposed clustering methods to a publicly available time course gene expression (microarray) data set from human subjects who were challenged by the influenza virus (GEO ID: GSE30550) (Woods et al., 2013). In this study, a cohort of 17 healthy human volunteers received intranasal inoculation of influenza H3N2/Wisconsin virus. Nine of the 17 subjects developed mild to severe symptoms. Following from Linel et al. (Linel, Wu, Deng, & Wu, 2014), we identified the top ranking genes (TRG's) that have the largest dynamic response to the influenza virus for each of the subjects. In this analysis, we will focus on the nine symptomatic subjects since we are interested in identifying early signals of clinical outcomes.

The proposed clustering approaches identified inhomogeneous clusters with different sizes, shapes and densities, namely large, medium, small and single-gene clusters. These four types of temporal gene response modules assume different roles in modulating the dynamic response to the disease. For each subject we identified temporal gene response modules that can be used to predict the severity of the influenza infection. We also discover a power-law distribution for the size of the temporal gene response modules, which indicates that the response of the underlying biological system is driven by a few universal characteristics, this phenomenon is referred to as universality in complex systems (Barzel & Barabási, 2013). These novel findings may help us to understand the redundant design at the genetic level of a biological system.

The remainder of the paper is organized as follows. Section 1 reviews the MCL and GMM algorithms and describes the proposed correlation-based iterative hierarchical clustering (IHC) and iterative pairwise-correlation clustering (IPC) methods in detail. In Section 3, we provide a comparative analysis of clustering results from the real data and computer simulation studies for the proposed methods, the MCL and GMM algorithms. We also present the biological findings related to the