

Bayesian analysis of cytomegalovirus infection from microarray data

Un'analisi Bayesiana di infezioni da cytomegalovirus da dati di tipo microarray

Debora Slanzi¹

Dipartimento di Statistica, Università Ca' Foscari, e-mail: debora@unive.it

Marta Trevisan

Dipartimento di Istologia, Microbiologia e Biotecnologie Mediche,
Università degli Studi di Padova, e-mail: marta.trevisan@unipd.it

Laura Villanova

Dipartimento di Scienze Statistiche, Università degli Studi di Padova,
e-mail: laura.villanova@stat.unipd.it

Giorgio Palù

Dipartimento di Istologia, Microbiologia e Biotecnologie Mediche,
Università degli Studi di Padova, e-mail: giorgio.palu@unipd.it

Irene Poli

Dipartimento di Statistica, Università Ca' Foscari,
e European Center for Living Technology, e-mail: irenepoli@unive.it

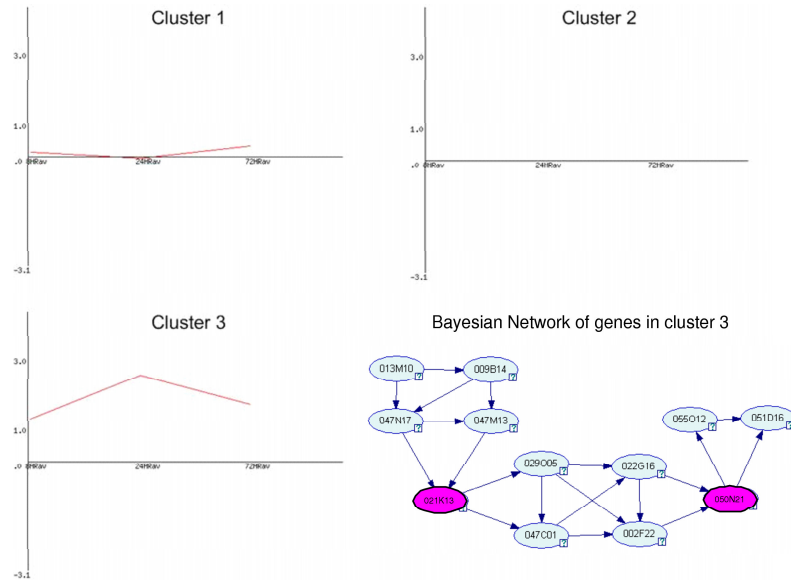
Riassunto: Le infezioni da Cytomegalovirus sono molto frequenti e sono tipo asintomatico; ma in soggetti immunocompressi tali infezioni possono degenerare in gravi malattie. In questo studio abbiamo analizzato il profilo dell'espressione genica ottenuto da esperimenti condotti con tecnologia cDNA microarray. Adottando un approccio Bayesiano nella derivazione di cluster di profili osservati, abbiamo individuato un piccolo gruppo di geni con livelli di espressione alti nelle prime ore dell'infezione. La struttura di dipendenza all'interno di questo cluster, individuata con una rete Bayesiana, evidenzia la presenza di due geni, geni digi, con un ruolo di controllo.

Keywords: Microarray gene expression data, Bayesian cluster analysis

Human Cytomegalovirus (HCMV) is a DNA virus belonging to the herpesvirus family. The frequency of infection is very high ($\sim 90\%$) in the general adult population and after a primary infection the virus enters a state of latency persisting lifelong. In the healthy population the primary infection is asymptomatic, whereas in immunocompromised, the infection leads to severe diseases. To gain insights into mechanisms of this infection we employed cDNA microarray technology to assess global changes in cellular gene expression (Speed, 2003). In this research, a class of adrenal cells were infected with HCMV; mock infection was performed by treating cells in the same way as those infected except for the absence of the virus. At different intervals of time after infection, 8h, 24h and 72h, cells were harvested for RNA purification and microarray analysis. RNA was then hybridized to cDNA microarray slides containing 21329 spotted oligonucleotide sequences of human genes. Microarray experiments were performed in duplicate including dye swap. In this paper we adopt a Bayesian approach for model based-clustering of gene expression profiles (Ramoni et al., 2002). We assume that the data we observe are generated

¹Address of correspondence: Dipartimento di Statistica, San Giobbe, Cannaregio 873, Venezia

Figure 1: Profiles of each cluster and the Bayesian network



by a set of unknown processes and then we build a clustering procedure to identify these processes. Under the hypothesis that two gene profiles are similar if they are generated by the same process, we consider a set of c clusters of gene expression profiles as a statistical clustering model M_c , consisting of C_k groups where every cluster groups k genes. In this framework, we adopt, as a similarity measure, the posterior probability of the clustering model. Initiating the procedure with the assumption that all clustering models are equally probable, we observe the data and compute the posterior probability of each model. The procedure is accomplished by using a similarity-driven search method. Goodness of fit of the resulting model is assessed by checking the normality of the standardized residuals of each cluster. The method is implemented in a computer program called CAGED (<http://genomethods.org/caged>). The study yields three clusters of gene expression profile as described in Figure 1. The cluster 3 is composed of only 12 genes, which are particularly up-regulated during the first hour post infection. Most of these genes are involved in angiogenesis or control of the cell cycle progression, others are responsible for calcium within the cell. To get more insight of the problem and learn on the dependence structure between these genes, we build a Bayesian Gaussian Network (Bottecher and Dethlefsen, 2003). As it is described in Figure 1, there are two crucial genes in the structure: 021K13 and 050N21, which play the role of dam genes controlling the modulation of the others. These results seem very interesting and we are currently investigating the role of these two genes in mediating the effects HCMV infection of the adrenal cells.

References

- Speed T.P.(2003) *Statistical Analysis of Gene Expression Microarray Data*. Chapman & Hall/CRC.
- Ramoni M.F., Sebastiani P. , Kohane I.S. (2002) Cluster analysis of gene expression dynamics. *PNAS*, 99(4), 9121-9126.
- Bottecher S.G., Dethlefsen C. (2003) Deal: a Package for learning Bayesian Networks. *Journal of Statistical Software*, 8(20).