BIOINFORMATICS LETTER TO THE EDITOR Vol. 24 no. 7 2008, page 1033
doi:10.1093/bioinformatics/btn019

Gene expression

Reply to 'Comment on causality and pathway search in microarray time series experiment'

Nitai D. Mukhopadhyay[1,*] and Snigdhansu Chatterjee[2]

[1]Department of Biostatistics, Virginia Commonwealth University and [2]Department of Statistics, University of Minnesota

Received on December 3, 2006; revised on January 7, 2008; accepted on January 10, 2008

Advance Access publication February 22, 2008

Associate Editor: Martin Bishop

We thank Professors Nagarajan and Upreti for their interest in our paper, Mukhopadhyay and Chatterjee (2007). There, we propose using Granger causality-based pathway detection in an acyclic, homoscedastic framework for microarray time-series expressions; which are generally short-duration time series involving very large number of genes. Professors Nagarajan and Upreti point out that in the presence of heteroscedasticity, and a cycle like 'gene x regulates the expression of gene y and simultaneously gene y regulates the expression of gene x', Granger causality tests may not be informative. Here, we adopt the term 'heteroscedasticity' ('homoscedasticity') to mean the unconditional variance of the white noise, represented as a bivariate vector in the Euclidean co-ordinate system, is different (same) in different co-ordinate directions.

Thus, in essence, if the assumptions about the acyclic and homoscedastic nature of the time series are violated, tests for causality detection may fail. This is an important point, since when a contemporaneous cyclic relationship is present, the notion of causality makes little sense. In the context of economics, Eichler (2007) present a treatment of contemporaneous correlation as well as Granger causality. Extreme heteroscedasticity may be indicative of improper normalization of gene expressions. At the end of their letter, Dr Nagarajan and Dr Upreti mention the normalization step. Proper normalization should remove wide discrepancy in noise variance, hence nowadays microarray datasets are typically available in *de facto* normalized version. The data used in Mukhopadhyay and Chatterjee (2007) is also normalized. However, difference in technical variance, as indicated by Professors Nagarajan and Upreti, may still be present. And that will violate the assumption of our method (as well as many other statistical comparison methods relying on common unknown variance).

Professor Nagarajan, in review, kindly suggested references for two-gene systems whose time-profile may not fit into to a homoscedastic, cause-effect framework. Thus, a full vector autoregression structure may be needed to capture their mutual dependence at various lags (including lag zero). It can be guessed that multi-gene systems exist whose temporal co-dependency nature is extremely complex. Although current knowledge about gene regulatory networks is limited, some biology experts we consulted believe that cyclical patterns may be found in large multi-gene networks as a part of a feedback procedure, if they are studied over long enough time spans. A proper approach to elicit such patterns would be to conduct multivariate, possibly non-stationary, time-series analysis with all the genes over a long time horizon. This is not feasible currently, since present state-of-the-art microarray time series experiments are of short duration and typically involve very large number of genes. Hence, restricting the network to acyclic ones is, in our opinion, a small price to pay to produce informative analysis. Future microarray experiments over longer duration, along with discoveries of biological and chemical properties relating to gene and protein interactions, will no doubt lead to better understanding of gene networks.

We would like to point out in Model 1 (Equation 2), α_{12}, α_{21}, σ_ε^2 and σ_η^2 need to be *known constants* for the mathematical displays (4)–(7) to hold. As they stand, displays (4)–(7) are missing the $O(n^{-1})$ terms with each estimated parameters if some (or all) of $\{\alpha_{12}, \alpha_{21}, \sigma_\varepsilon^2, \sigma_\eta^2\}$ are estimated from data, where n is the length of the time series data. Also, the equation for s_1 does not account for the fact that as univariate time series, both x_t and y_t are $AR(2)$ (autoregressive of order 2) process and not $AR(1)$. Similar comments hold for Model 2 (Equation 11). The difficulty of modeling microarray time-series can be appreciated from the fact that in the human cell cycle data considered in Mukhopadhyay and Chatterjee (2007), n was 12 in one experiment, while the time-series itself was 802 dimensional.

REFERENCE

Eichler,M. (2007) Granger causality and path diagrams for multivariate time series. *J. Econometrics*, **137**, 334–353.

*To whom correspondence should be addressed.

