Generalized Correlation Coefficient for Non-Parametric Analysis of Microarray Time-Course Data

Tan, Qihua; Thomassen, Mads; Burton, Mark; Mose, Kristian Fredløv; Andersen, Klaus Ejner; Hjelmborg, Jacob v. B.; Kruse, Torben A

Published in:
Journal of Integrative Bioinformatics

DOI:
10.1515/jib-2017-0011

Publication date:
2017

Document version:
Final published version

Qihua Tan[1,2] / Mads Thomassen[1] / Mark Burton[1] / Kristian Fredløv Mose[3] / Klaus Ejner Andersen[3,4,5] / Jacob Hjelmborg[2] / Torben Kruse[1]

Generalized Correlation Coefficient for Non-Parametric Analysis of Microarray Time-Course Data

[1] Unit of Human Genetics, Department of Clinical Research, University of Southern Denmark, 5000 Odense C, Denmark, E-mail: qtan@health.sdu.dk
[2] Epidemiology, Biostatistics, and Biodemography, Department of Public Health, University of Southern Denmark, J.B. Winsløws Vej 9B, DK-5000, Odense C, Denmark, E-mail: qtan@health.sdu.dk
[3] Department of Dermatology and Allergy Centre, Odense University Hospital, University of Southern Denmark, 5000 Odense C, Denmark
[4] Dermatological Investigations Scandinavia, J.B. Winsløwsvej 9, 5000 Odense C, Denmark
[5] Centre for Innovative Medical Technology, Institute of Clinical Research, University of Southern Denmark, 5000 Odense C, Denmark

Abstract:
Modeling complex time-course patterns is a challenging issue in microarray study due to complex gene expression patterns in response to the time-course experiment. We introduce the generalized correlation coefficient and propose a combinatory approach for detecting, testing and clustering the heterogeneous time-course gene expression patterns. Application of the method identified nonlinear time-course patterns in high agreement with parametric analysis. We conclude that the non-parametric nature in the generalized correlation analysis could be an useful and efficient tool for analyzing microarray time-course data and for exploring the complex relationships in the omics data for studying their association with disease and health.

Keywords: time-course, gene expression microarray, generalized correlation coefficient
DOI: 10.1515/jib-2017-0011
Received: March 14, 2017; **Revised:** March 31, 2017; **Accepted:** April 4, 2017

1 Introduction

The time course design is important in microarray studies in exploring global transcriptional responses to treatment or to biochemical stimulations during *in vivo* or *in vitro* experiments. The complex dynamic gene expression patterns across thousands of genes over the time-course experiment impose a challenging issue for identifying and statistical testing in bioinformatics and biostatistics [1], [2], [3], [4], [5], [6]. Moreover, because the microarray experiment measures expression levels for thousands of genes simultaneously [7] and due to heterogeneity in the regulatory reaction among the large number of genes which cannot be predefined, it is impossible to inspect the observed and the fitted time-course patterns for determining a proper parametric form for the model for each gene on the array.

Statistical modeling for dealing with nonlinear patterns can be complicated [8] and requires intensive computation in case of high dimensional data such as microarray or genome sequence data, where there can be diverse patterns of dependence not limited to linearity. By introducing fractional polynomials to a growth curve model, Tan et al. [9] proposed an automated procedure to capture the various time-dependent expression patterns in microarray gene expression study. For each gene, the procedure compares the performances among fractional polynomial models with power terms from a set of fixed values that offer a wide range of curve shapes and suggests a best fitting model. Although the integration of growth curves with fractional polynomials provides a flexible way to model different time-course patterns, the selection of best fitting model requires intensive computation for estimating various models in a predefined parameter space. As a result, the identification of significant time-course patterns is highly computer intensive and time consuming.

Qihua Tan is the corresponding author.

Recently, the generalized correlation coefficients have been frequently discussed [10] and their application to large scale genomic data illustrated through microarray gene expression data analysis [11]. We propose a combinatory approach that makes use of the generalized correlation coefficient, for rapid identification, evaluation and classification of heterogeneous time-course patterns in microarray studies.

2 Methods

Figure 1 depicts the workflow for our combinatory approach. We follow the illustration to introduce the steps in the analysis.

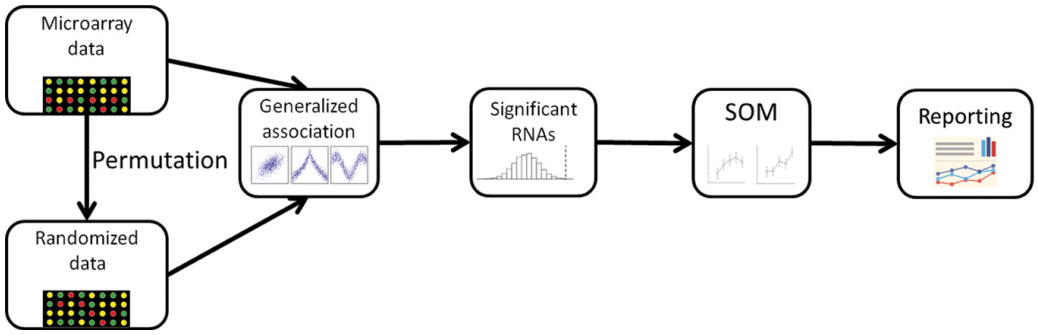

Figure 1: Flowchart of the combinatory approach for non-parametric analysis of microarray time-course data including steps starting from time-course detection, testing, clustering to final reporting.

2.1 Generalized Measures of Dependency

The generalized measures of dependency belong to the concepts of rank correlation and information theory based measures. The rank based correlation is well represented by Hoeffding's D [12] which measures the difference between the joint ranks of two random variables (X, Y) and the product of their marginal ranks. The information theory based approaches include mutual information (MI) [13] and maximal information coefficient (MIC) [11], [14]. By providing the amount of information one variable reveals about another, MI measures the dependency between two variables of any type. Based on the concept of rank correlation, Murrell et al. [15] very recently proposed a generalized correlation coefficient for non-parametric measurement of association between variables. The association score, A, ranges from 0 (when the variables are independent) to 1 (when they are perfectly associated). A is a kind of the square of the correlation coefficient that can be thought of as the proportion of variance in one variable explained by another variable or by a number of other variables. Since the explained variance is 1 minus the unexplained proportion, it can be expressed as $1 - \sigma^2_{error}/\sigma^2_{total}$ where σ^2_{total} and σ^2_{error} are the average squared deviations from a flat "null" model and a deterministic "alternative" model respectively. Under the normal assumption in least squares regression, the squared deviations can be expressed as probability density so that $R^2 = 1 - \prod_i \left(\frac{P(x_i, y_i|\text{null})}{P(x_i, y_i|\text{alternative})} \right)^{2/n}$. The last part of the formula (the proportion of unexplained variance) is the geometric mean of the squared ratio of the probability of observing a data point under the null model over the probability of that data point under the alternative model [15]. Here, the formulation of correlation depends only on the ratio of the probability density between two models which do not necessarily require normality assumption. A generalized version of R^2, or the so called association score A, can be calculated as long as the probability distributions for the null and alternative models can be evaluated using a kernel density approach as described by Murrell et al. [15]. The generalization enables direct estimation of the association score A when the form of correlation is unknown. The generalized association approach has been shown to have more power than MIC and fast convergence [15]. We apply the generalized association to microarray time-course data to detect the dynamic patterns of gene expression over the experiment time and across genes without any assumption for the heterogeneous relationship.

2.2 Identifying Significant Genes

In order to assess statistical significance of the time course patterns for each gene, we introduce a computer permutation procedure to generate random samples to estimate a null distribution of random association score

A. Significance of the time-course pattern for a gene can be determined by referring the observed score to the null distribution. Here we combine statistical testing with correction for multiple comparisons in microarray study by estimating the family-wise error rate (FWER). For each permutation sample, we record the highest A score from all genes to generate a distribution of N maximum of all A scores in a permutation from N permuted random samples. FWER for a gene is then calculated as the proportion of the maximum random A scores in the N permutations higher than a given gene's A score in the original data. Because the maximum A score from each permutation is based on all genes tested, the FWER automatically controls for multiple testing in the microarray study.

2.3 The Self-Organizing Map (SOM)

The SOM converts complex, nonlinear statistical relationships between high-dimensional data items into simple geometric relationships on a low-dimensional display [16]. As such SOM is an effective tool for visualizing high-dimensional data and for displaying similarities among data. Here, we make use of the nice features of SOM for clustering and displaying the various time-course patterns identified and rendered as significant by the permutation test. By tentatively adjusting the x and y dimensions of the output map, the number of clusters of time-course patterns can be determined for final reporting.

All analyses in this study were performed using the free R software (https://www.r-project.org). The generalized association was done using R package *matie* (https://cran.r-project.org/web/packages/matie/) and SOM was conducted using *som* package (https://cran.r-project.org/web/packages/som).

2.4 Application

In a recent microarray study, gene expression patterns in response to contact allergy was investigated using a time-course design consisting of a model allergen, diphenylcyclopropenone (DPCP), repeatedly applied on healthy skin on the inner upper-arms of 10 healthy volunteers [17]. Each individual received a series of repeated challenges at 4-weekly intervals. The biopsies taken from DPCP reactions in the initial elicitation challenge and repeated challenges were used for examining gene expression patterns in response to time-course experiment. Expression levels were measured from 38,500 genes using the Affymetrix Human Genome U133 plus 2.0 array. The raw expression data have been deposited at NCBI's Gene Expression Omnibus with GEO Series accession number GSE71996.

Following the steps in Figure 1, we first estimated A scores for each of the genes and then performed 10,000 permutations for testing the statistical significance. A total of 1631 genes were found to display significant time-course patterns over the experiment period (FWER<0.05). Next, we applied SOM for grouping and visualizing the different time-course patterns in the significant genes. In Figure 2, we grouped all the significant time-course curves into eight clusters. SOM shows that significant genes predominantly belong to two patterns, the four clusters to the left side (clusters 1, 2, 5, 6) of 964 genes and the two clusters to the right (clusters 4 and 8) of 500 genes. The first pattern is characterized by down-regulation of gene expression with most of the genes exhibiting a rapid decrease in expression from time points 2–3, i.e. 4 weeks after first challenge. The regulatory response is stabilized thereafter with only slight decreases driven by repeated challenges over time. Interestingly, the significant genes in clusters 4 and 8 manifest an opposite pattern featured by an increase in gene expression with the most rapid acceleration 4 weeks after first challenge followed by a stabilized and gradual increase. Both patterns indicate that there is an adaptive regulatory control mechanism that operates to prevent unrestricted reactivity. Clusters 3 and 7 represent some weak time-course patterns with fluctuated expression at time point three.

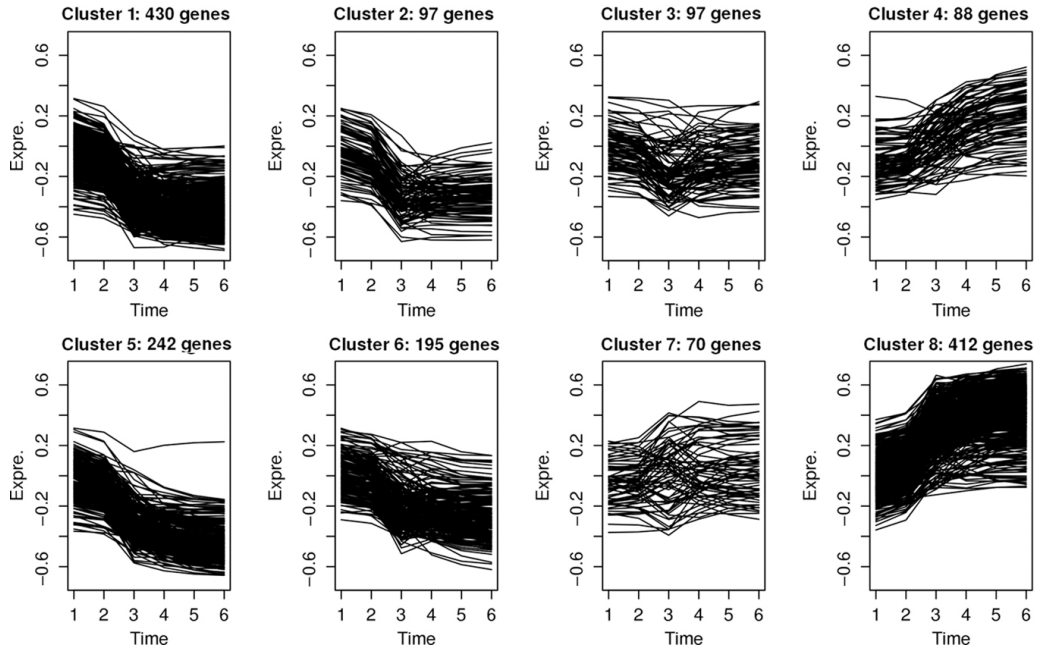

Figure 2: The heterogeneous time-course patterns for 1631 probe-sets identified by the generalized association analysis. The patterns are dominated by the down-regulation of gene expression in clusters 1, 2, 5 and 6 to the left and the up-regulation of gene expression in clusters 4 and 8 to the right.

3 Discussion

We have demonstrated a successful application of the generalized measurement of association to a microarray time-course experiment to detect significant genes manifesting heterogeneous regulatory expression patterns over the time of experiment. The analysis is characterized by (1) a non-parametric association that can capture any forms of response patterns induced by the time-course experiment; (2) permutation-based significance test that corrects for multiple testing in microarray study; and (3) focused clustering and visualization of significant time-course patterns for final reporting.

Our experience in parametric modelling of non-linear time-course data showed that computational load increases model complexity or the order of power function in case of fractional polynomials [9]. Even when computer time is not a factor of consideration, modelling complex patterns requires large sample sizes. Different from parametric modelling, the generalized correlation approach handles all patterns equally efficient and as such highly applicable to microarray studies which usually have limited scale of observations. Moreover, statistical analysis of high order polynomial models increases multiple testing and complexes significance assessment. In contrast, our permutation test applied to the non-parametric association deals with multiple testing in a simple and efficient manner.

In the analysis of Mose et al. [17], the application of the growth curve model with fractional polynomials identified 1556 probe-sets showing monotonous time-course patterns with FDR<0.05. Among the significant probe-sets, 871 overlap with the 1631 significant probe-sets from our generalized association analysis, an overlap rate of 56%. As our non-parametric analysis is not limited to monotonous patterns, we further compared our model performance with that of the parametric model. By fitting second-order fractional polynomials to the same data, 233 probe-sets were found to display non-monotonous time-course patterns with FDR<0.05. Among them, 181 are listed in the 1631 significant probe-sets in this study. That is 78% of the identified non-monotonous probe-sets are also found by our non-parametric association analysis. The high consistency in detecting non-monotonous patterns between parametric and non-parametric analyses suggests that the latter should be highly preferable for capturing the dynamic patterns in time-course microarray experiment especially when considering the complexity in modelling high order polynomial models.

In Figure 2, the fact that reactivity in gene expression reaches a plateau indicates that a regulatory control mechanism is operating that prevents an unrestricted increase in reactivity. As postulated by Mose et al. [17], one such inhibitory mechanism can be an inhibition of the continued expansion of relevant T effector cells, while another may be mediated by Tregs, which could suppress the activities of T effector cells through inhibitory cytokines, by cytolysis, or modulate the maturation and/or function of dendritic cells. Given the biological and clinical relevance of the identified time-course patterns, subsequent analysis should focus on functional annota-

tion to look for biological pathways or gene clusters linked to the time-course patterns to verify corresponding research hypothesis.

Recently, the nonparametric measure of association has been applied in feature selection for prediction model building in a proteomic study [18] and showed that, by relaxing the linear relationship assumption, the non-traditional method of association could help with more efficient feature selection while maintaining high prediction accuracy. The capability of handling both linear and nonlinear associations promotes the use of the generalized correlation coefficients in analysing massive and complex omics data with aim at ultimately disentangling and interpreting the complex patterns of relationships between omics data concepts in an integrative manner. Taking the relationship between gene expression and DNA methylation for example, multiple studies have been conducted with the purpose of analysing their correlation using Spearman's correlation coefficient. These studies have reported predominantly low or even poor correlation patterns [19], [20]. Here, we think that the more adequate generalized correlation methods should help to characterize the biological relationship more efficiently. Moreover, the generalized correlation can also be a useful tool for investigating the functional dependency between sets of attributes in omics data.

Acknowledgement

This work was financed by the Region Syddanmark research grant J.nr. 13/26014.

Conflict of interest statement: Authors state no conflict of interest. All authors have read the journal's Publication ethics and publication malpractice statement available at the journal's website and hereby confirm that they comply with all its parts applicable to the present scientific work.

References

[1] Storey JD, Xiao W, Leek JT, Tompkins RG, Davis RW. Significance analysis of time course microarray experiments. Proc Natl Acad Sci USA. 2005;102:12837–42.
[2] Bar-Joseph Z, Gerber G, Simon I, Gifford DK, Jaakkola TS. Comparing the continuous representation of time-series expression profiles to identify differentially expressed genes. Proc Natl Acad Sci U S A. 2003;100:10146–51.
[3] Lin T, Kaminski N, Bar-Joseph Z. Alignment and classification of time series gene expression in clinical studies. Bioinformatics. 2008;24:i147–55.
[4] Schliep A, Costa IG, Steinhoff C, Schönhuth A. Analyzing gene expression time-courses. IEEE/ACM Trans Comput Biol Bioinform. 2005;2:179–93.
[5] Costa IG, Schönhuth A, Hafemeister C, Schliep A. Constrained mixture estimation for analysis and robust classification of clinical time series. Bioinformatics. 2009;25:i6–14.
[6] Wichert S, Fokianos K, Strimmer K. Identifying periodically expressed transcripts in microarray time series data. Bioinformatics. 2004;20:5–20.
[7] Schulze A, Downward J. Navigating gene expression using microarrays–a technology review. Nat Cell Biol. 2001;3:E190–5.
[8] Royston P, Altman DG. Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. Appl Stat. 1994;43:429–67.
[9] Tan Q, Thomassen M, Hjelmborg JB, Clemmensen A, Andersen KE, Petersen TK, et al. A growth curve model with fractional polynomials for analysing incomplete time-course data in microarray gene expression studies. Adv Bioinform. 2011;261514.
[10] De Siqueira Santos S, Takahashi DY, Nakata A, Fujita A. A comparative study of statistical methods used to identify dependencies between gene expression signals. Brief Bioinform. 2014;15:906–918. DOI:10.1093/bib/bbt051.
[11] Reshef DN, Reshef YA, Finucane HK, Grossman SR, McVean G, Turnbaugh PJ, et al. Detecting novel associations in large data sets. Science. 2011;334:518–24.
[12] Hoeffding W. A non-parametric test of independence. Am Math Stat. 1948;19:546–57.
[13] Shannon CE, Weaver W. The mathematical theory of communication. Urbana, IL: University of Illinois Press, 1949.
[14] Speed T. A correlation for the 21st century. Science. 2011;334:1502–3.
[15] Murrell B, Murrell D, Murrell H. Discovering general multidimensional associations. PLoS One. 2016;11:e0151551.
[16] Kohonen T. Self-organizing maps, volume 30 of Springer Series in Information Sciences. Berlin, Heidelberg: Springer, 1995.
[17] Mose KF, Burton M, Thomassen M, Andersen F, Kruse TA, Tan Q, et al. The gene expression and immunohistochemical time-course of diphenylcyclopropenone induced contact allergy in healthy humans following repeated epicutaneous challenges. Experim Dermatol 2017. DOI:10.1111/exd.13345.
[18] Tan Q, Tepel M, Beck HC, Rasmussen LM, Hjelmborg J. Generalized measure of dependency for analysis of omics data. J Data Mining Genom Proteom. 2015;6:4.
[19] Wagner JR, Busche S, Ge B, Kwan T, Pastinen T, Blanchette M. The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. Genome Biol. 2014;15:R37.

[20] Kulis M, Heath S, Bibikova M, Queirós AC, Navarro A, Clot G, et al. Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia. Nat Genet. 2012;44:1236–42.