MEDINFO 2019: Health and Wellbeing e-Networks for All
L. Ohno-Machado and B. Séroussi (Eds.)

doi:10.3233/SHTI190559

A Search Method to Support Temporal Transcriptome Analysis

Guenter Tusch[a], Shahrzad Eslamian[b]

[a] *Health Informatics and Bioinformatics Program, School of Computing and Information Systems, Grand Valley State University, Allendale, MI, USA,*
[b] *Blue Cross Blue Shield of Michigan, Detroit, MI, USA*

Abstract

Recently, there has been an increasing interest in mining time series databases with a focus on data representation. We propose a hybrid statistical and temporal logic model to yield search pattern that can be used for pattern matching and database search to identify novel and common patterns in temporal expression experiments. The method accounts for various challenges that can be found in publicly available gene expression databases.

Keywords:

Gene Expression Profiling, Time, Database

Introduction

Public gene expression databases make enormous amounts of transcriptomic data available to researchers through the Internet. The most popular databases are GEO (Gene Expression Omnibus) at the NCBI and ArrayExpress at the EBI. NCBI GEO contains more than 105,000 series of gene-expression data collected from over 2.7 million samples and over 19,000 platforms (as of November 2018) and ArrayExpress over 2.3 million assays and some 71,500 experiments. Researchers attempt to translate the wealth of information captured in these rapidly growing biomedical and genomic databases into actionable data for health care or prevention, however the proper treatment of temporal data remains a challenge.

Biologists typically rely on statistical measures to determine trends. Time-course experiments allow them to study the dynamics of transcriptomic changes in cells exposed to different stimuli by investigating pathways or "curated gene sets". Those can be found in the Gene Ontology (http://www.geneontology.org), a framework defining concepts to describe gene function and their relationships, and the Kyoto Encyclopedia of Genes and Genomes (http://www.genome.jp/kegg/), a database integrating information on genomic and systemic function.

Traditional approaches to find gene pattern across studies are clustering and meta-analysis. Clustering approaches typically are based on GLM and splines to model profiles [1], while a meta-analysis attempts to either analyzing pooled data sets or aggregating p-values (see, e.g., [2]. One popular clustering approach for temporal data is statistical model-based clustering as describes for instance in [1] assumes that data is generated from a mixture model of probability distributions or spline approximations of the data. The mixture parameters can be estimated by maximum likelihood estimation.

A different approach used to model complex time pattern in patient records is Knowledge Based Temporal Abstraction (KBTA) to transform the raw data into a qualitative representation of temporal change based on time intervals, not discrete temporal data. KBTA is the task of summarizing large amounts of time-oriented data using domain-specific knowledge. For KBTA genomic clustering see [2].

The goal of our study is to enable the user to compare his/her own data to public datasets, regardless of the platform used.

Methods

Finding patterns of interest in time series databases (Query by Content) can be described as follows: Given query time series and some similarity measure, find the most similar time series in a given database. The two primary difficulties in this kind of similarity search are time complexity and defining a similarity measure.

Time complexity refers to efficiently finding a sub pattern in a larger pattern sequence. The approach in this paper is based on temporal logic. Finding sequential patterns in a database is implemented through Horn-like rules that can be expressed in terms of Web Ontology Language (OWL) concepts and that can reason about OWL individuals. A query is represented in a logical clause structure and a pattern matcher is used to search for complex sequence patterns of interest in a given database. This could include logical interdependencies between the elements of a sequential pattern including potential constraints as combining individual profiles in pathway or integrative genomics, e.g., genome-wide microarray expression signatures.

With our approach there are some similarities between clustering and database search. A database search for similar temporal pattern can be viewed as finding a temporal cluster that includes the query sequence. The key difference is here the variety of data sources. While clustering is typically applied to data from one experiment and one single platform, the queries under consideration here might include data from different, for example, Affymetrix platforms. While genes from different species can be compared using ortholog databases like PANTHER (pantherdb.org), the data (expression profiles) itself most likely won't fit the same statistical model due to different measurement techniques. The problem exacerbates if we compare profiles stemming from different manufacturers like Ilumina, or even microarray and next generation sequence data.

For our method, in order to increase the signal-to-noise ratio, the first step is signal averaging, i.e., averaging a set of replicate measurements of gene expression values. After averaging and piecewise linear approximation, temporal modeling through KBTA allows for conversion of expression values into an interval-based qualitative representation in a similar way as the well-known SAX method [4]. Change in the proposed method is determined by statistical significance of the paired t-test in consecutive time points. If the difference is significant, the interval is labeled as increasing or decreasing depending on the direction of change. Therefore, for the proposed method the p-values and the direction of the change inform the temporal abstraction. Time

points in temporal gene expression data are not standardized but change from experiment to experiment. If a biological signal is present, it is assumed not to depend on the interval length. Our method allows to compare studies independently from time points or scaling in the time axis or amplitude.

Since gene expression studies measure up to 40K genes or gene products on a single array with very few replicates two issues are critical for our implementation, first, to increase the power of the statistical tests (paired t-test) by empirical Bayes methods, and second, p-value adjustment. A moderated t-statistic shrinks the polled variance by borrowing information across all genes of the particular chip. Since a multitude of genes has to be tested for differential expression on the same data set, the p-value has to be adjusted. We use the False Discovery Rate.

The voom transformation RNA-seq data to be handled the same way as microarray data.

Implementation

Most of the data are preprocessed and stored in a MySQL database for performance reasons. Our plan is to migrate to an Apache Hadoop environment using Apache Hive when the numbers of the available time series data increases significantly. As of November 2018, we found in NCBI GEO 4357 microarray data sets (or 21,320 arrays) and 323 RNA-seq datasets (or 15,539 arrays). We use the GDS format with annotation files, which are normalized, and then extract the data matrix for microarray data. For high throughput sequencing RNA-seq studies data are preprocessed and normalized using standard procedures. We use R Bioconductor (BioC - bioconductor.org) with limma throughout and additionally the voom transformation for RNA-seq data. The necessary databases are accessed using standard BioC tools like GEOmetadb. This implementation has been integrated into the SPOT web application via HTML, JavaScript and PHP that feeds into Protégé. For modelling complex time patterns, we use Protégé (http://protege.stanford.edu) with the help of Allen's temporal logic as has been described in an earlier publication. We used as an example a microarray study with differentially expressed genes grouped into three peak clusters.

Discussion and Conclusions

Our model differs from statistical model-based clustering as it only assumes the general linear model, but does not model the time courses explicitly. This approach allows to cluster gene expression profiles that stem from different platforms, even microarray and nextgen sequencing RNA-seq data sets while still taking the temporal information into account. Our approach also borrows from KBTA. One of the weaknesses of the KBTA approach is that arbitrary ("knowledge based") thresholds have to be determined. In our application thresholds would need to be determined for each platform, manufacturer and technology. The key advantage of our approach is that it implicitly estimates the thresholds from the data itself and uses statistical significance to generate biologically meaningful clusters. The temporal relationship is represented by comparing the means in each pair of groups.

Due to the complexity of the transcriptome and the limited knowledge so far, there are two main approaches to evaluate the feasibility of our method. The first is to rely on existing across platform studies especially with known homologous genes and use the results and determine some form of recall and precision. The second is to perform a simulation study and generate artificial gene expression sets with realistic features. There are a variety of cautionary points: Besides limma, there are other common algorithms to determine significance in time-series gene expression data, as for instance SAM, EDGE or BETR. Typically, a p-value is assigned to an entire gene set or time series, often incorporating a sophisticated weighting schema [4-7] Our approach could be adapted.

If a microarray study or RNA-seq study finds significant gene expression changes, those typically are verified through fluorescent, one-step reverse transcription-polymerase chain reaction (RT-PCR) or quantitative polymerase chain reaction (qPCR). Since so far there are several viable alternative approaches (esp. for RNA-seq data), each study potentially uses a different method to identify differentially expressed genes and similarly for clustering the gene sets as for example STEM, GQL [4-7] or TimeClust [2], the significant genes that our algorithm finds, may not be verified. One solution could be implementing different standard pipelines and giving the user the choice. Given, that we describe an exploratory and not modelling approach, we face the same challenges as other search tools in bioinformatics in terms of how to implement biological knowledge in a computer system.

We will compare ours to those published results and continue to evaluate our program on a representative set of sample studies.

References

[1] D. Lu, Y. Tripodis, L.C. Gerstenfeld, and S. Demissie, Clustering of temporal gene expression data with mixtures of mixed effects models with a penalized likelihood. *Bioinformatics,* **35(5)** (2019) 778–786.

[2] P.H. Sudmant, M.S. Alexis, and C.B. Burge, Meta-analysis of RNA-seq expression data across species, tissues and studies. *Genome biology*, **16 (**2015) 287.

[2] L. Sacchi, R. Bellazzi, C. Larizza, P. Magni, T. Curk, U. Petrovic, and B. Zupan, TA-clustering: Cluster analysis of gene expression profiles through Temporal Abstractions. *International Journal of Medical Informatics* **74** (2005), 505-17.

[4] Y. Shahar, A framework for knowledge-based temporal abstraction. *Artificial intelligence* **90** (1997), 79-133.

[5] C.W. Law, Y. Chen, W. Shi, and G.K. Smyth. Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome biology* **15** (2014), R29.

[6] G. Tusch, O. Tole, and M.E. Hoinski, A Model for Cross-Platform Searches in Temporal Microarray Data. In *Conference on Artificial Intelligence in Medicine in Europe 2015 Jun 17*. Springer Internat. Publ. 2015. (pp. 153-158)

[7] F. Vázquez-Chona , B.K. Song, and E.E. Geisert Jr, Temporal changes in gene expression after injury in the rat retina. *Invest Ophthalmol Vis Sci* **45** (2004), 2737-46.

[8] Z. Bar-Joseph, A. Gitter, and I. Simon, Studying and modelling dynamic biological processes using time-series gene expression data. *Nature Reviews Genetics* **13** (2012), 552-64.

[9] L. Sacchi, C. Larizza, P. Magni, and R. Bellazzi, Precedence Temporal Networks to represent temporal relationships in gene expression data. *Journal of Biomedical Informatics* **40** (2007), 761-774.

Address for Correspondence

Dr. Guenter Tusch, One Campus Drive, D-2-238 MAK, Allendale, MI 49401, USA. Email: tuschg@gvsu.edu