A Model for Temporal Abstraction in Gene Expression Studies

Guenter Tusch[1] and Shahrzad Eslamian [1]

[1] Grand Valley State University, Allendale, MI, USA
tuschg@gvsu.edu, eslamias@mail.gvsu.edu

Abstract. When utilizing information from increasingly voluminous biomedical and genomic databases into actionable data for health care, treatment of temporal data still remains a challenge. Frequently temporal research is based on stimulus response studies and includes searching for temporal effects or time patterns in gene sets. Digital gene expression (DGE) technologies like rna-seq seem to replace microarray technologies in the near future for many functional genomics applications.
This study explores the feasibility of searching for temporal patterns based on knowledge-based temporal abstractions. Those imply conversion of expression values into an interval-based qualitative representation expressing amount of change over time. The amount of change is determined by statistical significance. For microarray studies one approach uses Bioconductor limma software modelling the normalized intensities in the framework of the linear model. Empirical Bayes methods result in a moderated t-statistic that reduces the pooled variance by borrowing information across all genes. We use the moderated paired t-test to determine significant differences in consecutive time points. While this approach assumes that the experiment is based on one particular platform, comparison across platforms can be done by comparing p-values. Therefore, in our model the p-values and the direction of the change inform the temporal abstraction. We discuss this approach in the framework of our SPOT software.

Keywords: Temporal representation and reasoning, statistical, decision support, microarray, rna-seq.

1 Introduction

When translating information captured in increasingly voluminous biomedical and genomic databases into actionable data for health care or prevention, treatment of temporal data still remains a challenge. The challenge is not so much in modeling complex pattern of intervals like in clinical domains (see [1]), but more in classifying different types of intervals in terms of trends, as “increasing”, “decreasing”, “constant” etc. (see [2]). Biologists typically rely on statistical measures for those purposes. Frequently temporal research is based on stimulus response studies and includes searching for temporal effects or time patterns in gene sets or pathways across data from different studies. A large body of information is available in public repositories like NCBI GEO and ArrayExpress. It appears that digital gene expression

(DGE) technologies like rna-seq will replace microarray technologies in the near future for many functional genomics applications.

This paper explores the feasibility of searching for temporal pattern based on temporal modeling through knowledge-based temporal abstractions that allow for conversion of expression values into an interval-based qualitative representation expressing the amount of change over time. It also allows to compare studies where the experimenter chose different pattern of time points. Change in these types of studies is typically determined by statistical significance. Assume a researcher conducted a temporal study where he/she discovered peaks in a set of genes that might be found in the same biological pathway. He/she now wants to see if finding the same effect in related studies can extend his/her hypothesis. Although different studies address similar questions a comparative search through public databases is impeded by the use of heterogeneous platforms and analysis methods. We describe a model that can help alleviate those problems.

2 Methodology

Databases that contain temporal gene expression data are organized in a way that data are recorded at the particular time point the measurement took place, i.e., the tissue sample was taken. These time points are not standardized but change from experiment to experiment and are determined by the experimenter as he/she sees best fit for the biological question that will be answered by the experiment. If we look for time patterns, e.g., peaks, in that database, standard query languages like SQL are not helpful here when searching across different experiments with potentially different time point patterns.

1.1 Temporal Abstraction

We use Knowledge Based Temporal Abstraction (KBTA) to transform the data into a qualitative representation of temporal change based on intervals, not discrete temporal data. KBTA is the task of summarizing large amounts of time-oriented data using domain-specific knowledge (see Shahar [2]). The KBTA method is based a formal model of input and output entities, their relations, and the domain-specific properties that are associated with these entities - called the KBTA ontology. Shahar describes four different output types, state, gradient, rate, and pattern abstraction. For the domain of gene expression studies predominantly the gradient type representing temporal trends is important. It might represent increasing, decreasing or constant values within a specific time interval and could be labeled “increase”, “decrease”, “constant”, etc. The temporal interval can form patterns using Allen’s approach of temporal relationships [4]. For example, a “peak” can be defined as an increasing interval immediately followed by a decreasing one. Thus peaks can be found even if experiments use different time point patterns.

There is a variety of implementations of the KBTA method in different domains, many clinical. Almost all of them focus on describing individual patient courses for therapeutic purposes, e.g. [1], but there are a few genomic applications, e.g., [3]. The

key to this methodology is to use domain-specific knowledge to determine, if values are changing or remaining constant. Several packages and tools are currently available, for instance the virtualArray software package in Bioconductor can combine raw data sets using almost any chip types based on current annotations from NCBI GEO. No such tool is currently available for temporal studies.

The goal of this study is to make temporal information available that can be found in publicly available repositories. We use as an example NCBI GEO that contains data from both microarray and rna-seq gene expression studies. For microarray studies, it contains in most of the cases both raw data sets and curated data. To accommodate for potential bias based in experimental conditions data typically have to be cleaned and normalized before they can be used for analysis. There are different normalization procedures, which are chosen by the authors of the publication to their best knowledge. The curated data sets (Genomic Data Structure - GDS - in NCBI terms) are normalized. Although this includes a subjective element, we chose to use GDS data whenever available, because the best represent the statistical results communicated in the corresponding publications. To implement KBTA in gene expression studies we use mostly the same methodology that a biologist would use to determine trends in high-throughput data, i.e., by means of statistical significance.

1.2 Domain Specific Knowledge

Due to the relative high cost of high-throughput sequencing technology sample sizes in gene expression studies are in general small resulting in little statistical power. To accommodate for that, empirical Bayes methods are employed. For microarray studies one approach uses the Bioconductor limma software modelling the normalized intensities in the framework of the linear model. Using the empirical Bayes methodology, a moderated t-statistic [5] is calculated that reduces the polled variance by borrowing information across all genes of the particular chip. We use the moderated paired t-test to determine, if there are significant differences in consecutive time points. If the difference is significant, we label the interval as increasing or decreasing depending on the direction of change. We don't adjust for the length of the interval assuming that, if a biological signal is present, it does not depend on the length of the interval. While this approach assumes that the experiment is based on one particular platform, comparison across platforms can be done by comparing p-values assuming that p-values accurately measure biological effects, and those don't depend on the platform. Therefore, in our model the p-values and the direction of the change inform the temporal abstraction, e.g., no significance means "constant".

1.3 RNA-seq Studies

It is expected that emerging digital gene expression (DGE) technologies will overtake microarray technologies in the near future for many functional genomics applications. In contrast to microarrays, rna-seq array require a computing intensive reassembly step for up to 300M reads with subsequent steps, typically using the Tuxedo protocol [6]. One of the fundamental data analysis tasks, especially for gene expression studies, involves determining whether there is evidence that counts for a transcript or exon are significantly different across experimental conditions [7]. There are at least

five different competing approaches for differential expression (DE) analysis, described by different software packages that implement those: ballgown, Cuffdiff2.1, EdgeR, DEseq2, DEGseq, BaySeq, Voom[9], etc. Some approaches are based on assuming a negative binomial distribution. For DGE count data we use the voom transformation [9]. Applied to the read counts it converts the counts to log-counts per million with associated precision weights allowing RNA-seq data to be analyzed the same way as microarray data. This allows us to have a unified approach and use basically the same program for MA and rna-seq data. The result of the process is again a gene/value matrix with associated moderated t-statistics as in the previous section. The empirical Bayes approach has been applied to count data as well under the negative binomial distribution and several above mentioned programs employ that approach. Future research will show which approach gives the most reliable and trustworthy results, although our chosen edgeR/limma/voom approach with TMM normalization seems to be a likely candidate (see, e.g., [8]).

1.4 P-value Adjustments

For DE studies many genes have to be tested for differential expression on the same data set, which leads to a depreciation of the the nominal p-value. Therefore, the p-value has to be adjusted. Two methods are typically applied, the Bonferroni correction or the false discovery rate (FDR) approach. Both approaches have their drawbacks and are missing some significant genes. While the focus of the FDR is on controlling the false positives while potentially missing many significant genes, the Bonferroni correction is very conservative in assigning significance. FDR is the most common approach; therefore, we use that adjustment also to have a more likely match with published results and biological verification. Significant genes are verified, e.g. by qPCR, in almost all publications.

Since both the Bonferroni correction and FDR are potentially fail to detect a few significant genes, researchers frequently apply prior knowledge about known highly correlated gene sets, for instance biological pathways as they are collected in the KEGG pathway database or groups of genes that have the same functional annotation in Gene Ontology (GO). One popular approach is Gene Set Enrichment Analysis that calculates a single p-value for an entire gene set.

3 Implementation

For performance reasons most of the data are preprocessed and stored in a MySQL database. For microarray studies we use the GDS format with annotation files, which are normalized, and then extract the data matrix. For high throughput sequencing RNA-seq studies data are preprocessed as described above and normalized resulting in a data matrix. We use R Bioconductor (BioC) with limma and the voom transformation for rna-seq. The necessary databases are accessed using standard BioC tools like GEOmetadb. This implementation has been integrated into the SPOT web application [10] via HTML, JavaScript and PHP. Complex time patterns can be modelled using Protégé as has been described in an earlier publication [10].

The above described approach works for searches within a species and if all platforms involved in the search share the same genes. For searches across species we determine the orthologs, i.e., genes in different species that evolved from a common ancestral gene by speciation, from the InParanoid database and translate genes in between platforms of different species.

4 Discussion and Future Aspects

While microarray DE studies are pretty much standardized, there is quite a variety of different pipelines for the analysis of rna-seq data sets. This poses a challenge for our model to treat rna-seq data since the results from our unified approach may differ from the actual confirmed results in the corresponding publications, because different analysis pipelines typically find different differentially expressed albeit mostly overlapping gene sets [8]. Since only gene sets from the original publication are potentially confirmed by qPCR, our tool might use unconfirmed results. One solution could be implementing different standard pipelines and giving the user the choice, which one to use. Given, that we describe an exploratory approach here, this might however be of minor concern. It is also not intended for modelling (see e.g. [11]).

References

1. Moskovitch R, Shahar Y. Classification-driven temporal discretization of multivariate time series. Data Mining and Knowledge Discovery. 2015 Jul 1;29(4):871-913.
2. Shahar Y, Musen MA. Knowledge-based temporal abstraction in clinical domains. Artif. Intell. Med. 8(3) (1996) 267-98.
3. Sacchi L, Larizza C, Magni P, Bellazzi R. Precedence temporal networks to represent temporal relationships in gene expression data. Journal of Biomedical Informatics. 2007 Dec 31;40(6):761-74.
4. Allen JF. Towards a general theory of action and time. Artif Intell 23(2) (1984) 123-154.
5. Smyth G. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Statistical Applications in Genetics and Molecular Biology, 3:Article 3, 2004.
6. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature protocols. 2012 Mar 1;7(3):562-78.
7. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szcześniak MW, Gaffney DJ, Elo LL, Zhang X, Mortazavi A. A survey of best practices for RNA-seq data analysis. Genome biology. 2016 Jan 26;17(1):13.
8. Seyednasrollah F, Laiho A, Elo LL. Comparison of software packages for detecting differential expression in RNA-seq studies. Briefings in bioinformatics. 2015 Jan 1;16(1):59-70.
9. Law CW, Chen Y, Shi W, Smyth GK. Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome biology. 2014 Feb 3;15(2):R29.
10. Tusch G, Tole O, Hoinski ME. A Model for Cross-Platform Searches in Temporal Microarray Data. In Conference on Artificial Intelligence in Medicine in Europe 2015 Jun 17 (pp. 153-158). Springer International Publishing.