Summary of Ph.D. thesis

Bioinformatic Analysis and Digital Signal Processing on Global Gene Expression Screening Data

János-Zsigmond Kelemen

Supervisor: László Puskás PhD, DSc

Laboratory of Functional Genomics

Biological Research Center of the

Hungarian Academy of Sciences

University of Szeged

2007

Introduction

The ever-increasing flow of biological data – DNA sequence, gene expression profiles, protein-protein interactions – leads to rapid progress in the area of biology known as systems biology. The available high-throughput gene-expression quantification technologies are partly responsible for the burst of this field. In an attempt to model and simulate the biological system of the cell, systems biology promises better understanding of life functions and also reliable treatment against disease. It is known that the various subtypes of cancer respond differently to various treatments. It is essential, therefore, to accurately diagnose a tumor, before any treatment. Based on its gene-expression profile, a tumor cell can be viewed as a state machine with each state corresponding to the biological state of cancer subtype. This leads to the idea of gene-expression based molecular classification - a mathematical approach to cancer diagnosis, which is a true systems biological task. This sort of class prediction problem, particularly based on DNA microarray data, has

been an important research topic in recent years. A large number of machine learning algorithms and methods, such as support vector machines, artificial neural networks, nearest neighbor classifiers, or random forests, have been applied, aiming for better accuracy and precision of diagnosis, and also the selection of a more reliable cancer signature consisting of a reduced number of genes. Unfortunately, the gene expression data used for such classifications is invariably corrupted with noise, either of biological or of experimental origin. Thus, for a reliable classification, the data has to flow through various preprocessing stages.

Objectives

The aims of this study are typically concerned with gene expression data processing. The list of objectives related to the subsequent individual bioinformatic processing steps is presented below.

- Application of the “gold standard” gene-expression data analysis methods to real laboratory QRT-PCR and microarray data.

- Statistical analysis of the effect of laboratory protocol innovation on the gene-expression experiment outcome.
- Class discovery and marker gene testing in schizophrenia transcriptional profiles.
- Development of innovative system level methods for expression data normalization and noise reduction (Kalman Filter), with application to molecular diagnosis of cancer.

Incorporating the expression covariance between genes proves to be an important issue in biological data classification problems with application to diagnosis, since this represents the functional relationships that govern tissue state. We also aim to show here that employing the Kalman Filter on microarray data to remove noise (while retaining meaningful covariance and thus being able to estimate the underlying biological state from microarray measurements) yields linearly separable data suitable for most classification algorithms.

Results

Since this dissertation is concerned with numerical processing methodologies for biological data, the results here are practical implementations of methods to real gene expression data. The actual biological results, although significant, were not of major concern here. The basic and compulsory data preprocessing steps, namely quality control and LOWESS normalization, and also the *t*-test for detecting the differentially expressed genes are exemplified using publications that I have coauthored. A detailed description of the actual implementation of these procedures is given for the experiment related to the identification of the genes modulated by the *ntrR* gene in *Sinorhizobium meliloti*. These methods or similar are applied however in all the experiments related to this study.

An experiment concerning the expression changes induced by lipopolysaccharide treatment on mouse macrophage cells was used to assess for the effect of the amplification protocol used for sample preparation, on the detected expression changes. A statistical analysis based

on the custom application of the $\chi 2$ test on the categorical expression change results (down-regulation, up-regulation, no change) for some 15 genes, shows that the exponential-phase DNA amplification is more reliable than the saturation-phase over-amplification for sample preparation. These results are important for selecting the proper protocol, from the reproducibility point of view.

A more complex analysis of transcription profiles is presented within an experiment seeking to identify genes regulated differently in schizophrenia compared to the healthy control. During the analysis, two genes, namely *DRD2* and *Kir2.3*, were identified as having such a behavior. These genes were proposed as marker genes. To test their predictive capability in diagnosing the disease, a hierarchical clustering was performed on data samples specific to these two genes, coming from both healthy and schizophrenic individuals. The unsupervised method discovered the two biologically distinct classes.

At the same level of analysis complexity, we were also concerned with classification (supervised clustering) of microarray data, as a molecular diagnosis method for

cancer subtypes. Here we proposed the Kalman filtering procedure as a mathematical tool which is able to decompose the noise into biologically meaningful variance and measurement noise or error.

Considering the biological state the true gene expression profile associated with a tumor family, the biological variance is the stochastic model of the expression changes associated with the tumor subclasses under investigation. The measurement noise, on the other hand, represents the stochastic model of all the errors that can appear at the various laboratory phases in the course of a microarray experiment. The Kalman filter, using a state-space model of the data flow, and the two mentioned stochastic models, estimates the actual biological state.

We applied Kalman filterning on seven publicly available cancer expression datasets, and tested the support vector machines, artificial neural networks, nearest neighbor classifiers, and random forests classification methods before and after filtering. In a mostly technical discussion of the Kalman filtering results with regard to classification, we show that the

classification results were significantly improved. Three state-of-the-art graphical representation schemes are also employed in the study, to inspect whether the tumor subclasses are also visually detectable. We also discuss in detail the selection of marker genes. The predictive potential with regard to cancer, of the original and Kalman filtered marker genes is assessed statistically, and we show that the number of Kalman filtered features necessary for a good discrimination of tumor types is smaller than the size of the raw feature set required for a similar performance.

Publications

Related to this thesis:

Kelemen JZ, Kertesz-Farkas A, Kocsor A, Puskas LG. Kalman filtering for disease-state estimation from microarray data. Bioinformatics. 2006 Dec 15;22(24):3047-53. Epub 2006 Oct 25.

Zvara A, Szekeres G, Janka Z, **Kelemen JZ**, Cimmer C, Santha M, Puskas LG. Over-expression of dopamine D2 receptor and inwardly rectifying potassium channel genes in drug-naive schizophrenic peripheral blood lymphocytes as potential diagnostic markers. Dis Markers. 2005;21(2):61-9.

Nagy ZB, **Kelemen JZ**, Feher LZ, Zvara A, Juhasz K, Puskas LG. Real-time polymerase chain reaction-based exponential sample amplification for microarray gene expression profiling. Anal Biochem. 2005 Feb 1;337(1):76-83.

Puskas LG, Nagy ZB, **Kelemen JZ**, Ruberg S, Bodogai M, Becker A, Dusha I. Wide-range transcriptional modulating effect of ntrR under microaerobiosis in Sinorhizobium meliloti. Mol Genet Genomics. 2004 Oct;272(3):275-89. Epub 2004 Sep 9.

Other publications:

Feher LZ, Balazs M, **Kelemen JZ**, Zvara A, Nemeth I, Varga-Orvos Z, Puskas LG. Improved DOP-PCR-based representational whole-genome amplification using quantitative real-time PCR. Diagn Mol Pathol. 2006 Mar;15(1):43-8. Erratum in: Diagn Mol Pathol. 2006 Jun;15(2):123.

Posters

Péter Hankovszky, Liliána Z. Fehér, **János Zsigmond Kelemen**, Sándor Sonkodi, László G. Puskás; Vasoactive effects of the long-term administration of erythropoietin and global gene-expression analysis in the rat mesenteric artery, 2005. Istanbul ERA-EDTA XLII Congress - poster

Bodogai M., Puskás L.G., Nagy Zs.B., **Kelemen J.Zs.**, Rüberg S., Becker A., Dusha I., Wide-range transcriptional modulating effect of ntrR under microaerobiosis in Sinorhizobium, The 6th European Conference on Nitrogen Fixation, 2004 - poster

Oral presentations

Kelemen JZ. "Methods in Microarray Analysis – Classification", Second International School on Biology, Computation and Information (BCI 2005), Dobbiaco (BZ), Italy, September 12-16, 2005