BIOINFORMATIC ANALYSIS AND DIGITAL SIGNAL PROCESSING ON GLOBAL GENE EXPRESSION SCREENING DATA

by

János-Zsigmond Kelemen

Supervisor:

László Puskás, PhD, DSc

A dissertation submitted in partial
fulfillment of the requirements for the
degree of

Doctor of Philosophy

University of Szeged

2007

To my mother and to Enikő

„My root was spread out by the waters, and the dew
lay all night upon my branch.
My glory was fresh in me, and my bow
was renewed in my hand."
Job 29,19-20 KJV

TABLE OF CONTENTS

ACKNOWLEDGMENTS

I wish to thank Dr. László Puskás - my supervisor and mentor - for giving me the opportunity to learn and work in his group. I am also grateful to him for giving me a chance to think and implement my own ideas as well. Secondly, I acknowledge the help of my colleagues at the Laboratory of Functional Genomics, BRC HAS. Special thanks to Dr. Kocsor András and Kertész-Farkas Attila (Research Group on Artificial Intelligence of HAS and University of Szeged), for their valuable collaboration on the Kalman filtering project. Last but not least, I thank my dear Enikő for her patience all those years, and for standing by me at all times.

ABBREVIATIONS

QRT-PCR. Quantitative real-time PCR.

MIAME. Minimum Information About a Microarray Experiment.

CP. Crossing points.

LOWESS. Locally Weighted Scatter-Plot Smoothing.

SVM. Support Vector Machine.

ANN. Artificial Neural Network.

kNN. k Nearest Neighbors algorithm.

RF. Random Forest.

ROC. Receiver Operator Characteristic.

RFE. Recursive Feature Elimination.

LLE. Locally Linear Embedding.

KF. Kalman Filter.

ARE. Algebraic Riccati Equation.

LPS. Lipopolysaccharide.

SRBCT. Small, round blue cell tumors of childhood.

EWS. The Ewing family of tumors.

BL. Burkitt lymphoma.

NB. Neuroblastoma.

RMS. Rhabdomyosarcoma.

LIST OF FIGURES AND TABLES

FOREWORD

This section is intended as plead for interdisciplinary research. For some obscure reason, Alfred Nobel[i], the father of the notorious homonym prize, decided to omit mathematics from among the distinguished sciences. Various rumors have surrounded the peculiar decision of the famous mecena. One of them states that Nobel decided against a prize in mathematics because his fiancé cheated on him with a famous mathematician, often claimed to be Gösta Mittag-Leffler[ii]. The most likely explanation is, however, that he considered mathematics a purely theoretical science with no direct practical benefit to mankind. If that was the case, modern research and the fusion of sciences that emerged thereof proved him wrong. Informatics, the latest branch of mathematics, including game theory, control theory, graph theory or algorithms for that matter, has been thoroughly integrated with other disciplines, being an indispensable tool of current research. Representative figures have demonstrated the practical importance of this area hitherto considered entirely theoretical. Let's just take John Nash[iii] for a well known example, since he was the hero of the Hollywood movie Beautiful Mind. But the choice of the 2002 Nobel laureates (to mention just two: Daniel Kahneman[iv] and Vernon Smith[v] – economy, both owing much to game theory) also shows the role mathematics plays in all the sciences. Thus, Mittag-Leffler is cheating again on Nobel by taking the prize the back-door way. Putting the anecdote aside, today's science has become profoundly computer-centric. Not only the huge libraries have become available by means of informatics, but also the modern tools of research. The future of understanding nature, as many see it, lays in the concept of interdisciplinarity. Perhaps the "homo universalis" idea of the Renaissance has failed, but in order to understand our environment and life itself, as in putting the pieces together, we certainly recognize the need for a

modern "universal" research group, where "Les savants ne sont pas curieux" (Anatole France[vi]) does not apply.

When I was graduating the university, one of my professors held a speech. His closing remark was that all he expects us, graduate engineers, to remain with after five years of training, is a systemic thinking. Thus, faithful to this idea I shall continue and during this thesis I shall present a gradual evolution and development of biological data analysis from simpler bioinformatic and statistical analysis to a systemic signal processing framework. First, let us introduce systems theory as an interdisciplinary theory that is concerned with the properties of systems as wholes. As opposed to studying the individual system components, systems theory studies the interactions between these components, interactions that will determine the general properties and behavior of the system. Established as a science by Ludwig von Bertalanffy[vii], Anatol Rapoport[viii] and others in the 1950s, systems theory can be considered a revolutionary change of the scientific view of the world. The major practical applications of this field are found in control engineering. Currently, however, the emerging scientific discipline of systems biology is also beginning to use the achievements of systems theory. Thus, systems biology, often overlapping with bioinformatics, integrates molecular biology knowledge and computational analysis in an attempt to model, simulate and analyze biological systems and processes.

AIMS

The aims of this study are typically concerned with gene expression data processing. The list of aims related to the subsequent individual bioinformatics processing steps, as illustrated by our results, 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 to remove noise on gene expression data (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.

Chapter 1

GENERAL INTRODUCTION

1.1 Mathematical Modeling - A Short Survey

"The mathematics is not there till we put it there" (Arthur Eddington[ix]), therefore, a mathematical model is not the system but an abstract model that uses mathematical language to approximately describe the system as in Figure 1.

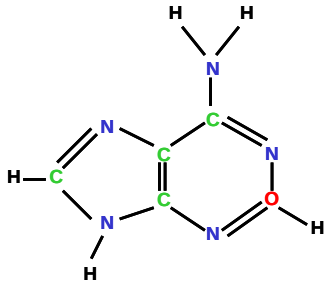

Figure 1. Graph model of the adenine nucleotide integrating physics and chemistry knowledge and possibly predicting biological function.

There is a close relationship between mathematical modeling and systems theory. To make it more clear, a mathematical model is a representation of the properties of a system in a mathematically usable form [1]. The components of the system are represented by variables, while the relationships between them by mathematical functions. There are three major objectives of system modeling: analysis of system structure, prediction of behavior and ultimately control of behavior. These objectives bear with major consequences; let's just think of the biological system of the cell for example. In the following we will shortly describe the essential aspects of mathematical modeling and clear some concepts that will be used later on throughout this dissertation. Models can be static or dynamic. Static models, such as those used for classification in this study (random forests,

k-nearest neighbours) do not account for the element of time. They can be linear (linear kernel support vector machines) or non-linear (sigmoid output artificial neural networks), depending on the nature of the functions relating the variables, and they are trained in order to learn the structure of the system that produced the so called training dataset (for example a microarray dataset that accounts for several disease subclasses). The training itself can be supervised or unsupervised, depending on whether the user will supervise the training by providing the expected output for the given input data (classification, also known as supervised clustering), or the model should self-organize in order to fit the input data (hierarchical clustering). Dynamic models, on the other hand, attempt to model the time dependent relationships between the variables (state-space models). Here we can have output variables (actual measurements), input variables (system control) or state variables (hidden variables such as the true gene expression state that we attempted to estimate using the Kalman filtering). Models can also be deterministic, such as a state-space model that uses the state transition functions to uniquely determine the following state from the current one, or stochastic. Stochastic models use probabilistic approaches to account for the randomness of the variables (for example the noise models used by the Kalman filter are stochastic). The evaluation of an acquired model is of particular importance, since it concerns the model's reliability (classification models for medical diagnosis). For this purpose, a set of test data is usually used. If the model shows comparable performance on the training data and on the test data, then the model fits well the system in cause. However, there is a degree of uncertainty when it comes to the model handling events outside the measured data. Eddington has a witty solution to that problem: "It is also a good rule not to put overmuch confidence in the observational results that are put forward until they are confirmed by theory".

1.2 Biology of Gene Expression

The DNA molecule encodes the hereditable information fundamental to the life of the cell. A major discovery of molecular biology was that the DNA encoded biological information is copied by the RNA and that the RNA mediated information is used to assemble the proteins. Proteins thus decode biological information into biological function. This flow of information (Figure 2) from DNA to RNA and from RNA to protein is stated as the "Central Dogma" of molecular biology, which was proposed by Francis Crick[x] in 1957 [2].

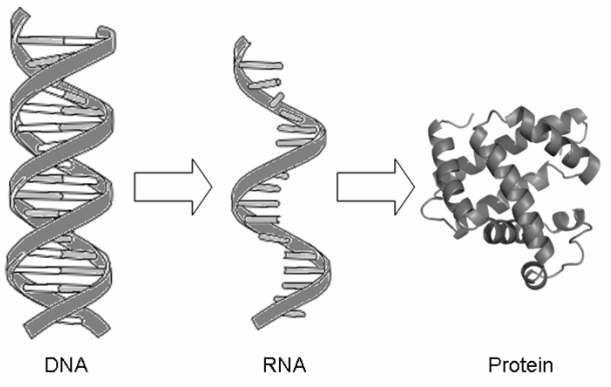

Figure 2. Flow of information: in transcription DNA information is copied to produce an RNA transcript; in translation the instructions in mRNA are used to synthesize a polypeptide.

The actual mechanism of gene expression is complex and consists of two major stages. During transcription, the transcript of the gene is produced as a molecule of mRNA. During the second phase, namely translation, the mRNA nucleotide information is decoded to a sequence of amino acids yielding the polypeptide at the ribosomes. The transcription of genes is a complexly regulated process. A large network of signal mediating components (signal pathways) is involved in activating the final effectors of the process, such as the transcription factors or the RNA polymerase [2][3]. Identifying this so called transcriptional network represents one of the major goals of systems biology. The mathematical modeling

of such a large system can be made possible by the currently available high throughput technologies that provide large scale data like gene sequence, transcription profiles, protein quantitative data as well as protein interaction data. Clearly, having such a model would have major implications in predicting the response of the cell to stimuli for example, or in controlling the cell's behavior. For the time being however, based on the modeling the available biological data, the less ambitious task of medical diagnosis can be achieved, still of great importance in medicine.

1.3 Biotechnological Research Tools

As we saw, according to the "Central Dogma", the gene transcript stands in the information flow path between the DNA and the protein. Thus, quantifying the mRNA provides in some degree quantitative information about the proteins downstream, and also qualitative information about the DNA upstream. It has been shown that transcription profiles reflect well the biological state of the investigated samples. Thus, measuring the mRNA level at a particular cell state also provides insight into gene expression events, genes being activated and partly proteins that the cell responds with, allowing us to infer their function. Currently two major technologies are available for gene transcript quantification: quantitative real-time PCR (QRT-PCR) and DNA microarray. The RNA level measured using these techniques actually corresponds to the stationary level of the RNA formed as a combination of transcription, RNA maturation and RNA degradation.

1.3.1 Real-time PCR

The "real-time" or kinetic variant of the polymerase chain reaction invented by Kary Mullis[xi], was pioneered by Higuchi et al. [4]. The DNA is quantified after each amplification cycle by detecting the fluorescence emitted by a dye

intercalating with the double-strand DNA. For mRNA quantification, reverse transcription PCR (RT-PCR) is used first to reverse transcribe the RNA to complementary DNA (cDNA), which is then quantified using the QRT-PCR.

1.3.2 DNA Microarray

In recent years, a new technique, the DNA microarray technology (DNA-chip) [5] has emerged, offering the possibility of high-throughput systematic analysis of the transcriptome. The arrays are constructed of thousands of DNA fragments either spotted or synthesized (Affymetrix) onto chemically activated glass slides. DNA fragments can be collections of short or long oligonucleotides or cDNAs of variable length. DNA microarrays with sets of cDNA fragments on their surface can be used to obtain a molecular fingerprint of gene expression of cells. The method has enabled large numbers of genes, from specific cell populations, to be studied in a single experiment.

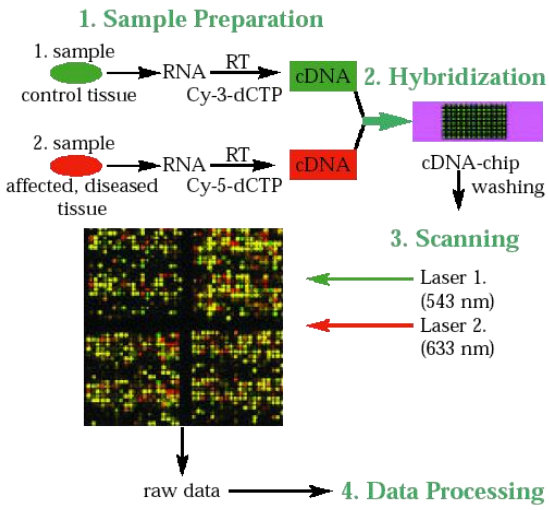

Figure 3. Application of cDNA arrays for the follow up detection of gene expression changes.

During the experiment, mRNA populations gained from diverse biological samples (tissues or cell cultures) are converted to cDNA in the presence of fluorescent dye (Cy3 or Cy5) labeled nucleotides. Using a co-hybridization strategy, with Cy3- labeled cDNA from the test sample and Cy5-labeled cDNA from the control sample, the relative intensity ratio on the microarrays can be determined and the expression pattern can be analyzed. The schematic representation of a cDNA microarray experiment can be seen in Figure 3. Clearly the most attractive feature of the technology and its major advantage over the QRT-PCR is the high throughput.

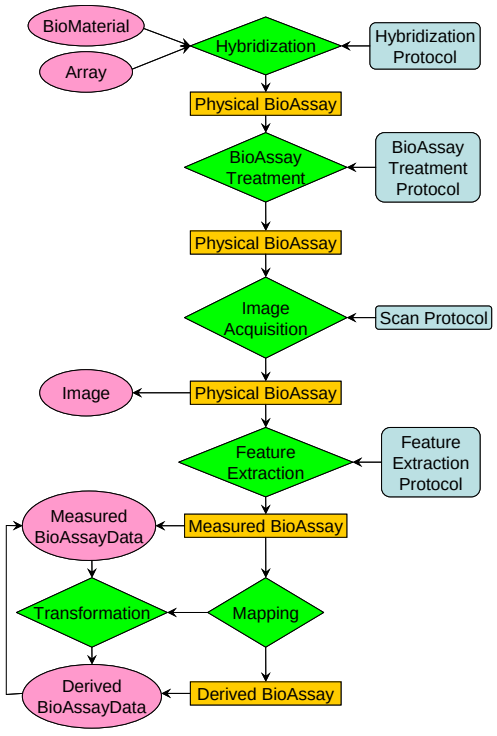

Figure 4. Block diagram of the microarray experiment taken from the MIAME exchange specifications regarding the microarray workflow.

This however induces some disadvantages as well, namely the high number of error sources which will act in the detriment of precision. As opposed to the

QRT-PCR, the DNA microarray is static, based on the natural pairing of the complementary bases. The dynamics of the PCR facilitate the more exact analysis of the differences in gene expression. It is a good practice, therefore, to verify the results of a microarray experiment by means of real-time PCR. Currently, the primary applications of microarrays include gene discovery [7], disease diagnosis [8], drug discovery [9], and toxicological research [10]. More advanced systems biological tasks, such as transcriptional network modeling, have also been attempted [11][12][13]. A successful DNA microarray experiment starts with a good design of the experiment. Figure 4 shows such a design and specific workflow steps according to the MIAME specifications [6]. The block diagram of the microarray experiment reveals the complexity and the large number of stages involved in the process. It has to be outlined that each of the blocks corresponding to the work phases, including hybridization, washing (BioAssay treatment), image acquisition, and the protocols used, are major sources of error and noise that will be superimposed on the final data output. Due to this noise, subsequent data pre-processing and bioinformatic analysis steps must be taken before any biological interpretation of the results.

Chapter 2

METHODS AND MODELS FOR GENE EXPRESSION DATA ANALYSIS

The methods described below aim at data obtained as result of quantification by means of either arrayed image analysis (microarray) or change in fluorescence signal intensity (during the PCR reaction). These digitized signals are considered raw measurement data and reach the analysis workstation in the form of numerical vector columns. Within such a vector each observation element corresponds to the expression of a specifically inspected gene under the given conditions. They will be called alternatively features, while the vectors themselves will sometimes be denominated as samples.

2.1 Quality Control and Feature Rejection

Quality control [14] can refer to any step that is required to prepare a generic data set for a specific type of analysis. Quality control is often referred to as data filtering, a term that can take different meanings as we shall see later on. The process primarily involves two concepts: feature rejection and averaging. The various quality control measures provided by the quantification process are used to eliminate specific observations that do not comply with given laboratory set or standard thresholds. Flags set at the image analysis stage, local background estimates, signal to noise ratios are valuable information in determining the reliability of features in a microarray measurement. The same applies for QRT-PCR data, where dilution curves, reaction efficiencies and crossing point deviations can motivate the rejection of certain reactions. Technical replicas are repeats of measurements of samples coming from the same pool of extracted RNA. They are truly meant to control the quality and reproducibility of the experimental conditions. Averaging is generally used to combine observations

from measurements technically replicated. Biological replicates on the other hand, are independent measurements of the same type. They provide the means for most of the methods presented in the following sections. Quality control is a concept that can be applied at any stage of processing and it is a compromise between biological and mathematical consideration, since there is a general danger of losing data before one is certain that it is not useful or unusable. The practical application of quality control within the case studies shown in the Results section was done in MS Excel.

2.2 Normalization – A Milestone

Normalization at its origins denotes a transformation of the data, which results in a normal (Gaussian) distribution. The denomination here has historical reasons since the parametric statistics used for differentially expressed gene identification require normal distributions. There is general agreement that a log transformation of most microarray data provides a good approximation of the normal distribution with minor exceptions [15]. Currently, however, the normalization procedure covers more tasks essential to expression data analysis including scaling, noise smoothing and dye bias correction. To conclude, the gene expression data related normalization represents the procedures required to make the samples comparable. There is no optimal general method to be used, thus, depending on the experiment various normalization schemes can be employed.

2.2.1 Reference features

Using control features to normalize the expression data is a popular method, which is based on the assumption that certain genes do not change their expression under the inspected circumstances. One must be careful however, since endogenous controls (i.e. housekeeping genes) can become unreliable under rough changes in the cell's housekeeping such as tumor. In the following we shall

describe one such normalization procedure widely and exclusively used with QRT-PCR data, known as the Pfaffl method [16]. This method provides a means for quantification of a target gene transcript in comparison to a reference gene. The relative expression ratio is calculated only from the real-time PCR efficiencies and the crossing point deviation of an unknown sample versus a control. The model used needs no calibration curve, as control levels are included within the model. High accuracy and reproducibility (less than 2.5% variation) may be reached using this procedure. For the mathematical model it is necessary to determine the crossing points (CP) for each transcript. CP is defined as the point at which the fluorescence rises appreciably above the background fluorescence. CP cycles versus cDNA concentration are then plotted to calculate the slope (mean, standard deviation). The corresponding real-time efficiencies are computed according to the equation:

$$E = 10^{-\frac{1}{\text{slope}}} \tag{1}$$

The relative expression ratio (R) of a target gene is calculated based on E and the CP deviation of an unknown sample versus a control, and expressed in comparison to a reference gene:

$$R = \frac{E_{\text{target}}^{\Delta CP_{\text{target}}(control-sample)}}{E_{\text{ref}}^{\Delta CP_{\text{ref}}(control-sample)}} \tag{2}$$

The ratio of a target gene is expressed in a sample versus a control in comparison to a reference gene. E_{target} is the real-time PCR efficiency of target gene transcript; E_{ref} is the real-time PCR efficiency of a reference gene transcript; $\Delta CP_{\text{target}}$ is the CP deviation of control – sample of the target gene transcript; ΔCP_{ref} is the CP deviation of control – sample of reference gene transcript. The reference gene should be a stable and secure unregulated transcript. Because the perfect

reference does not exist, reference genes should be validated by showing that they do not change significantly in expression under the experimental conditions. Using multiple reference genes can increase reproducibility. Examples of common internal standards include b-Actin, GAPDH, MCH I mRNA, and ribosomal RNAs (rRNA). For microarray data, where thousands of genes are monitored, things tend to be more difficult. The literature is acquainted with the housekeeping genes method [18] and the "spiking" technique [17][19], although there is no standard on how the actual normalization should be performed using the reference genes. The application of the Pfaffl method to real laboratory QRT-PCR data within our projects was done in MS Excel.

2.2.2 Global methods

Most published references to microarray normalization deal primarily with the removal of biases in the data. Bias arises from a number of sources, including variation within and among arrays, differences in mRNA concentration or quality, unequal dye incorporation, and wavelength-related differences in scanner strength. Without correcting these biases, it may appear as though too many genes are up- (or down-) regulated. Bias correction is performed based on some assumption that the experimenter makes. The first is that the starting amount of cDNA used for each hybridization is the same. This type of assumption can also be made in the case of QRT-PCR. The preferred method to use in these cases is the so called total intensity method [20], a global normalization procedure. The technique consists of a simple scaling usually so that the sums across the samples (i.e. the overall sample intensities) become equal. Other variants include mean or median centralization. To present the method in a mathematical form, first we introduce the following model of hybridization for a single spot on the array:

$$T_i = aC_i + b_i \tag{3}$$

where T_i represents the fluorescence intensity of the i-th test spot and C_i is ibid for control. The b_i parameter stands for the difference in expression and is considered a random variable, normally distributed around null:

$$\mu(b_i) \approx 0 \Rightarrow \sum_i b_i \to 0 \tag{4}$$

We are interested in the parameter a, the constant normalization factor across the data samples. To express the equal starting cDNA amount assumption we write:

$$\sum_i T_i = \sum_i (aC_i + b_i) \tag{5}$$

From the two above equations the scaling factor is derived as:

$$a = \frac{\sum_i T_i}{\sum_i C_i} \tag{6}$$

2.2.3 Regression methods

Global centralization cannot correct for biases that are present within specific parts of the data, mostly due to unequal dye incorporation or spatial irregularities on the physical array. For these systematic errors, presuming we can make the assumption that the great majority of genes do not change their expression within the experiment, regression methods are the choice. These approaches are particularly important when using ratios to monitor changes in gene expression and especially when employing a two-color scheme. To visually identify such bias problems a graphical aid such as the M vs. A plot can be used [21]. The measured expression (M) is the logarithmic gene expression ratio between the test and the control samples. The intensity (A) represents the average of log intensities over the test-control cases. Thus, for each spot i we have:

$$M_i = \log_2 \frac{T_i}{C_i}$$
$$A_i = \frac{\log_2 T_i + \log_2 C_i}{2} \quad (7)$$

With these transformations we can again write the hybridization model as follows:

$$M_i = a_i + b_i \quad (8)$$

where this time the scaling factor is not a constant, but an intensity dependent value for each feature; log-expression change b_i again follows a Gaussian distribution:

$$a_i = f(A_i)$$
$$\mu(b_i) \approx 0 \quad (9)$$

The normalized expression log-ratios are then computed:

$$M_i^{norm} = M_i - f(A_i) = b_i \quad (10)$$

$f(A_i)$ is typically a regression function. The use of the locally weighted scatter-plot smoothing (LOWESS) [22] has been suggested [21][23] to correct the intensity dependent measurement corruption, being one of the most robust curve fitting procedures. Figure 5 shows LOWESS in action on an example dataset. As an effect of the procedure, the data "cloud" is smoothed, and centered around zero. The normalized dataset is shown in Figure 6.

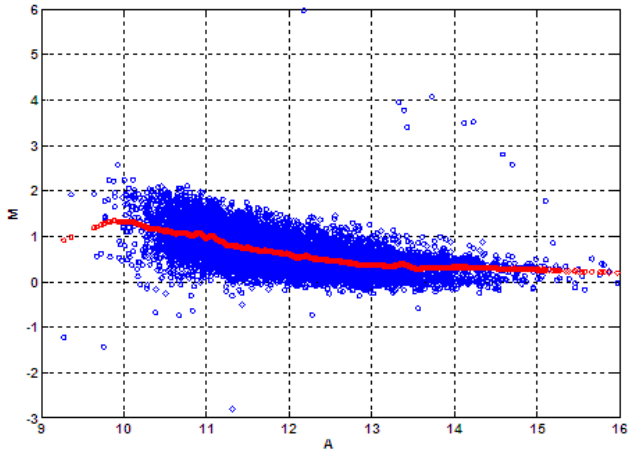

Figure 5. M vs. A plot of the raw dataset and the LOWESS curve.

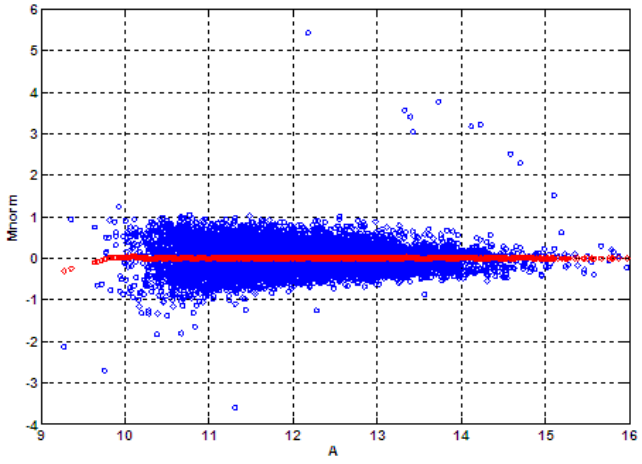

Figure 6. Normalized expression ratios with the corresponding LOWESS curve.

The LOWESS normalization within our projects, exemplified as results, was performed using the R statistical software. Finally we mention that the normalization problem is still an open one. Each method has its advantages and

drawbacks, and new approaches with sound biological and mathematical basis are always welcome.

2.3 Decision and Statistical Significance

Having performed the two previously mentioned steps, namely quality control and normalization, we practically end up with the fold changes between control and test conditions for each gene under investigation. Many of these values, however, are false changes, mainly due to the experimental errors. To assess experimental error, one basically needs to repeat the experiment and measure the variation. If both control and test are biologically replicated, hypothesis testing can be used to decide whether the expression of a particular gene is significantly different between the two conditions.

2.3.1 Hypothesis testing

Statistically speaking, expression change can generally never be verified, but only disproved. Thus, we typically have a null (no expression change) hypothesis and an alternative hypothesis, contradictory to the former. The "change" hypothesis is supported if we can show that there is evidence against the null hypothesis. Hypothesis testing [24] consists of three steps:

1. Setting up the null hypothesis H_0 and the alternative hypothesis H_1.

2. Using a test statistic to compare the observed values with the values predicted by H_0.

3. Defining of a region for the test statistic for which H_0 is rejected in favor of H_1.

The probability that H_0 is true given the observed test statistic is called the *p*-value of the test. The level of significance *a* of a test is the probability that the test

statistic falls in the rejection region if H_0 is true. The test statistic is usually largely influenced by the population sample size. Thus the expression change decision is risky when the number of biological repeats is small, which can lead to type I (false negatives) or type II (false positives) errors.

2.3.1 Student's[xii] *t*-test

A frequent parametric test statistic for expression change inspection [23] is the *t*-statistic. Student's *t*-test [24] assumes the normality of the distributions of the data involved. Thus log-transformed gene expression data are suitable for such an analysis. Having the test-control log-ratio data prepared, the simplest and straightforward approach to detecting the differentially expressed genes is the single sample *t*-test. This variant of the statistic compares the mean of a sample population with a given value. The null hypothesis is therefore that the mean of expression log-ratio values for a gene is null, that is, its expression remains the same in both test and control conditions. The expression of the *t*-statistic in this case is the following:

$$t = \frac{\overline{R}\sqrt{n}}{s_R} \tag{11}$$

where $\overline{R}$ and s_R are the estimated mean and standard deviation of the log-ratios respectively, while n is the number of repeated measurements. The *t*-value follows a *t*-distribution with $df=n-1$ degrees of freedom. In situations where the measurements are not paired, or two conditions relative to normal are to be compared, the unpaired two-sample t-test can be used. The statistic on this case has the form:

$$t = \frac{\overline{R}_1 - \overline{R}_2}{s_{12}} \tag{12}$$

Considering unequal sample sizes n_1 and n_2, the estimated standard error of the mean difference is:

$$s_{12} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \tag{13}$$

and the *df* parameter equals n_1+n_2-1. If we perform multiple tests in parallel, usual for microarray data, the level of significance for the whole set of tests does not equal the level of significance for the single tests. The simplest adjustment is the Bonferroni correction. The overall significance level a_a is derived from the significance level a of m single tests by

$$\alpha_a = \frac{\alpha}{m} \tag{14}$$

The same procedure can be applied for the adjustment of the *p*-values. The *t*-test and Bonferroni correction on our own data was performed using MS Excel.

2.3.2 χ^2-test

Suppose we have an experimental factor, whose optimization can improve either cost or precision of a microarray experiment. We are interested in statistically assessing the effect of this factor on the actual expression changes as well as directly on their significance. For this purpose we propose using the χ^2-test [24][25]. In the simplest case we can have two categories of experiments, *C1* and *C2*, given the factor. Within each experiment category the *t*-test can be used to infer the test-control gene expression changes at the chosen significance level. Thus, based on the *p*-values and *a*, the continuous expression ratios can be discretized, each gene receiving a categorical value of down-regulated (*dr*), up-regulated (*ur*) or not-regulated (*nr*). This data altogether may be presented in a 3 *x* 2 contingency table with 3 rows and 2 columns such as Table 1.

Row categories	***C1*** 1	***C2*** 2	Total
dr	f_{11}	f_{12}	S_{dr}
ur	f_{21}	f_{22}	S_{ur}
nr	f_{31}	f_{32}	S_{nr}
Total	S_{C1}	S_{C2}	n

Table 1. Contingency table.

The entries in the table are frequencies; each cell contains the number of genes in a particular row and a particular column. Thus, we deal with two factors: the experimental factor and the expression change factor. The null hypothesis is that there is no association between the two factors, equivalent to the statement that tampering with the experimental factor does not influence the expression changes. Next we calculate the frequency that we expect in each cell of the contingency table if the null hypothesis is true. The expected frequency in a particular cell is the product of the relevant row total and relevant column total, divided by the overall total. In a final step, we calculate the test statistic that focuses on the discrepancy between the observed and expected frequencies in every cell of the table:

$$\chi^2 = \sum \frac{(O-E)^2}{E} \tag{15}$$

where O and E are the observed and expected frequencies in each cell of the table. This test statistic follows the χ^2 distribution with degrees of freedom equal to *(rows-1)x(columns-1)*. If the overall discrepancy is large, then it is unlikely the null hypothesis is true. The application of the χ^2-test on the practical case of assessing the effect of the amplification protocol on the expression change outcomes, shown in the Results, was done using the R statistical software package.

2.4 Advanced Analysis – Multivariate Statistics and Biomimicry Models

In the following we shall focus on gene expression data containing several biological repeats of possibly heterogeneous samples. The considerably large amount of data will be treated as an *nxm* matrix, each of the *n* rows corresponding to the investigated genes, while the *m* columns stand for the actual sample measurements. It is common with microarray data to have a much larger dimension *n* than samples *m*. The analysis of such high dimensional data requires "intelligent" approaches, and the methods employed have mostly immigrated from the field of artificial intelligence. Currently a somewhat paradoxal cycling of information between branches of biology (ecology, population genetics, or physiology) and mathematical modeling, back and forth, can be observed. Models developed initially to mimic biological phenomena, such as artificial neural networks, are being reused to handle the large amounts of newly produced biological data. As molecular biology processes are being understood, it is probable that the knowledge therein will yield more "intelligent" models. The methods introduced in the following aim at what is called clustering in statistics. The discovery of the subset- or class-membership of the samples in a dataset in this manner can be either self-motive (un-supervised) or supervised, and the result can be interpreted as the discrete states the biological system (cell) may be in.

2.4.1 Hierarchical dendrogram models

Hierarchical clustering is perhaps the best-known clustering method for expression data analyses. The main objective of this technique is to produce a tree like structure in which the nodes represent subsets of an expression data set. Thus, expression samples are joined to form groups, which are further joined until a single hierarchical tree (also known as dendrogram) is produced. Several studies on the molecular classification of cancers and biological modeling have

been based on this type of algorithms [26]. Hierarchical clustering of microarray data has been particularly fruitful in cancer diagnosis [27], investigation of cancer tumorigenesis mechanisms [28], or identification of cancer subtypes [29]. There are different versions of hierarchical clustering, which depend on the metric used to assess the separation between clusters, the cluster merging direction or the merging method. The most commonly used metrics are the Euclidean distance (a distance metric):

$$\mathrm{d}(x, y) = \sqrt{\sum_i (x_i - y_i)^2} \tag{16}$$

where x and y are two sample vectors and i spans the entire gene space; and the Pearson correlation coefficient (a similarity metric). Concerning the merging direction, the method can be divisive (top-down), which starts with one large cluster that contains all data points, and splits off a cluster at each step, or agglomerative. In the agglomerative method (bottom-up), illustrated in Figure 7, each data point initially forms a cluster, and the two "closest" clusters are merged in each step.

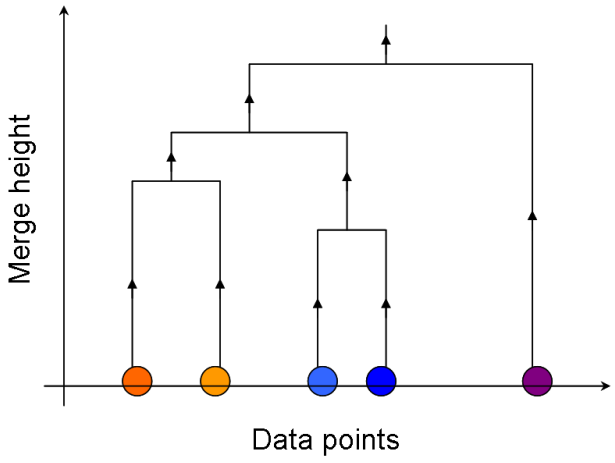

Figure 7. The process of agglomerative clustering. The distance between the data points is also suggested by the color spectrum.

The merging method or linkage can also be of various types including single linkage, average linkage and complete linkage. In single linkage clustering, the distance between any two clusters of points is defined as the smallest distance between any point in the first cluster and any point in the second cluster. Complete linkage defines the inter-cluster distance as the largest distance between any point in the first cluster and any point in the second cluster. Average linkage is often perceived as a compromise between single and complete linkage because it uses the average of all pair-wise distances between points in the first cluster and points in the second cluster. Thus, with average linkage, the distance between two clusters A and B is computed by:

$$\mathrm{D}(A, B) = \frac{1}{\mathrm{card}(A)\,\mathrm{card}(B)} \sum_{x \in A} \sum_{y \in B} \mathrm{d}(x, y) \tag{17}$$

The basic algorithm in hierarchical agglomerative clustering, using Euclidean distance and average linkage, is the following:

1. begin with each data point as a separate cluster;
2. using average linkage, merge the two clusters that are closest according to the Euclidean distance;
3. if only a single cluster remains then proceed with step 4, else redo step 2;
4. determine the final set of clusters.

The traditional approach for determining the final set of clusters is to specify the number of clusters desired and then cut the dendrogram at the height, which yields this number. In the schizophrenia gene expression profiling project we applied the hierarchical agglomerative clustering, with average linkage. The software we used for it was the R statistical environment.

2.4.2 Class prediction models

The following methods will particularly concern microarray data. It is known that a reliable and precise classification of tumors is essential for successful diagnosis and treatment of cancer. The gene expression-based molecular classification of cancer subtypes has been shown to have the potential of reliable diagnosis, either by complementing the traditional clinical, morphological and histo-pathological approaches or as an alternative procedure [29].

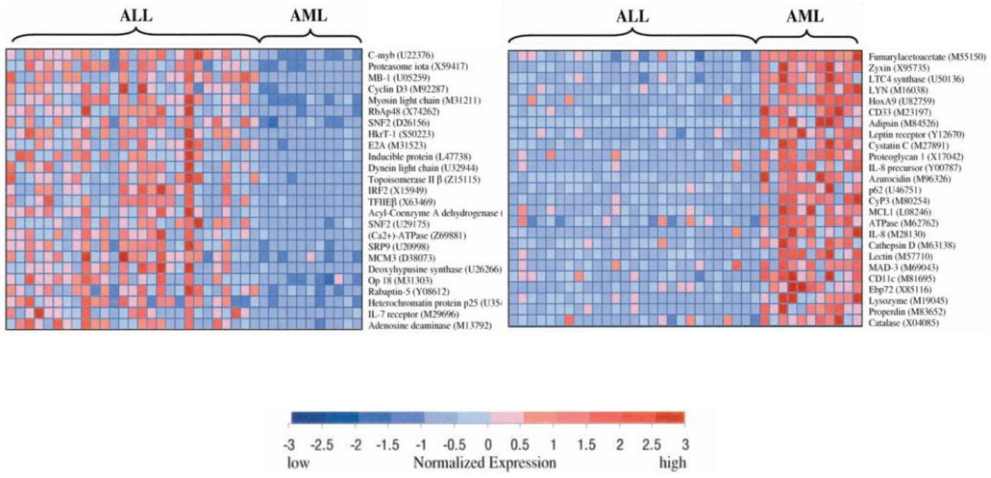

Figure 8. An example of gene expression based molecular classification of leukemia subtypes. Samples of acute lymphoblastic leukemia and acute myeloid leukemia were diagnosed.

The basic scheme of molecular classification is to train a mathematical model so that it can discriminate between the classes within a set of points. The training points in the expression data case are a set of gene expression measurements (e.g. microarray) that are fully annotated with regard to disease. In this case the genes represent the dimensions of the sample points. In the final classification or diagnosis phase, the model will automatically diagnose any new-coming test sample. An milestone example of such a classification based on gene expression profiles, as published by Golub et al [29], is presented in Figure 8. However,

having large datasets comprising simultaneous expression levels of thousands of genes monitored under diverse circumstances still constitutes a great challenge for biologists, physicians as well as computational algorithm developers. In recent years the processing of high-throughput biological data has evolved into a highly interdisciplinary field and a large number of machine learning algorithms have been proposed to automate difficult tasks, such as that of medical diagnosis from gene expression profiles. The following shortly reviews the most renowned of these algorithms and models, as they were employed in bioinformatics in general and in microarray data classification in particular.

The *Support Vector Machine* (SVM) classifier [30] is one of the most popular supervised learning algorithms, which has been effectively used in computational biology including protein remote homology detection [31], microarray gene expression analysis [32], the recognition of translation start sites [33], functional classification of promoter regions, the prediction of protein–protein interactions and peptide identification from mass spectrometry data [34]. The SVM classifier computes a hyper-plane with the largest margin between two classes [30] as seen in Figure 9.

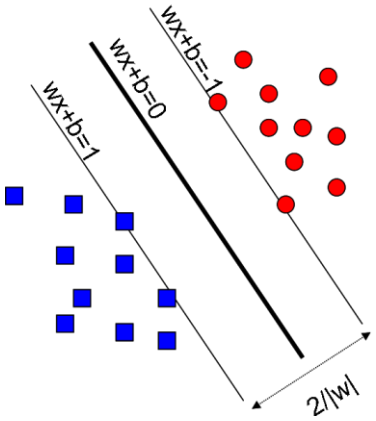

Figure 9. Maximum-margin hyper-planes separating the two classes within a training dataset.

Let us consider the training set of i high-dimensional points x, where x_i corresponds to the expression measurements of the ith experiment or sample. The a priori known two classes can be expressed by a label y_i associated with each x_i, such that $y_i \in \{-1,+1\}$. Assuming the classification function is linear, the label of a point can be written $y_i = \text{sign}(wx_i + b)$, where w is the normal vector to the hyper-plane separating the two classes, b is a free threshold parameter that translates the optimal hyper-plane relative to the origin, and operation wx_i is a dot-product. The distance from the hyper-plane to the closest points of the two classes is called the margin and is $\|w\|^{-2}$. The objective is to maximize the margin, with the constraint that the points from the two classes fall on opposite sides of the hyper-plane, written as:

$$\min_{w,b} \frac{1}{2}\|w\|^2 \text{ , subject to } y_i(wx_i + b) \geq 1. \quad (18)$$

This quadratic programming optimization problem is solved in its dual representation, which reveals that the classification is only a function of the support vectors, i.e., the training data that lie on the margin. In our experiments the SVMLight software [65] implemented in Matlab was used with a linear kernel.

The *Artificial Neural Networks* (ANNs) approach was originally developed with the aim of modelling information processing and learning in the brain [35][36][37]. Within the bioinformatics area this supervised nonlinear learner has been employed for instance in biological sequence analysis, the recognition of signal peptide cleavage sites, gene recognition [38], the prediction of protein functional domains [39] and the classification of cancer subtypes [40]. The ANN classifier consists of connected artificial neurons built in a multi-layer structure [35]. Thus, the basic unit of the neural network is the linear perceptron. As shown in Figure 10, the perceptron has n inputs x_i, $i=1:n$, and a single output y.

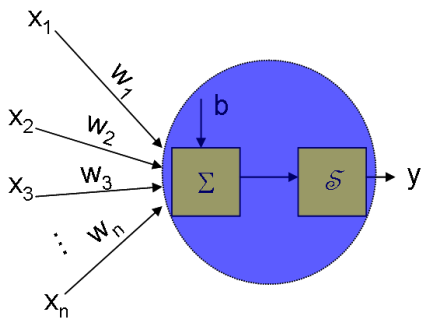

Figure 10. The artificial neuron or perceptron having *n* inputs and a single output.

Associated with each input is a weight w_i, that decides how important that input is for the output. To obtain the output, the weighted sum of the inputs, together with a bias b, is passed through an activation function S:

$$y = S\left(\sum_i w_i x_i + b\right). \tag{19}$$

The activation function is usually nonlinear, except for the input layer of the network. A typical activation function is the logistic sigmoid function:

$$S(z) = \frac{1}{1+e^{-z}}. \tag{20}$$

The single perceptron is a linear classifier similar to a linear SVM.

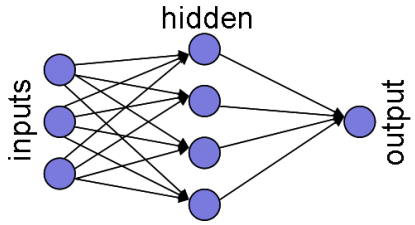

Figure 11. Multilayer perceptron with one hidden layer between the input and output layers.

Organizing the linear perceptrons in layers, as in Figure 11, results in a nonlinear classifier, which can effectively handle more difficult classification problems, such as multi-cancer diagnosis. Clearly the obtained neural network needs to be trained in a supervised fashion, using a train dataset and a set of class specific label values. In our study related to the Kalman filtering, a three layer ANN was used and the number of sigmoid output neurons within the hidden layer was determined by testing. Empirically we found that the best results were obtained with 25 hidden neurons. The ANN was part of the WEKA software package [66].

The *Nearest-Neighbor* (1NN) algorithm [41][42] is a simple class prediction technique, which achieves high-performance without a priori assumptions. This method has been used for protein classification [43] as well as cancer diagnosis [27]. The 1NN classifier is a fast algorithm, which is based on simple distance calculations between vectors.

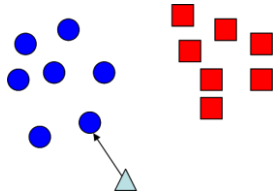

Figure 12. Classification criteria of the 1NN algorithm. The nearest point to the test case (triangle) is a circle, which determines the class membership.

Thus, the training phase consists only of storing the feature vectors and class labels of the training samples. In the actual classification phase, distances from the test cases to all stored vectors are computed and the closest sample is selected as pictured in Figure 12. The new point is predicted to belong to the closest class within the set. To measure the distance between gene expression samples, we used the Euclidean metric. The method can be easily extended to *k* neighbors (*k*NN). Within the Kalman filtering project, our tests showed that increasing *k* did not significantly improve the classification performance. The 1NN was typically outperformed by the previous two learners on the raw microarray data. The Matlab implementation of this algorithm was used in our study.

The *Random Forest* (RF) technique is a recently proposed meta-classifier method, which is becoming evermore popular in areas of computational biology like drug discovery [44] and tumor classification [45]. The RF technique is a combination of decision trees, such that each tree is grown on a bootstrap sample of the training set. For each node the split is chosen from a smaller subset of the total features, selected at random from an independent, identical distribution out of the feature set [46]. Thus, the method constructs a collection of decision trees with controlled variations. Let the number of training cases be *n*, and the number of features be *M*. Each tree is constructed using the following algorithm:

1. Select the number *m* of input variables ($m<<M$) to be used to determine the decision at a node of the tree.

2. Choose a training set for this tree by bootstraping. Use the rest of the cases to estimate the error of the tree, by predicting their classes.

3. For each node of the tree, randomly choose *m* variables (from an independent, identical distribution out of the feature set) on which to

base the decision at that node. Calculate the best split based on these *m* variables in the training set.

4. Each tree is fully grown and not pruned (as may be done in constructing a normal tree classifier).

The output of the RF is the class that is the mode of the classes output by the individual trees. In our experiments, 20 trees were used and *m* was set to *log(n + 1)*. The software that we used for it was part of the WEKA package [66].

For multi-class datasets the one-versus-rest technique was used. Thus, for every biological class an independent binary learner was built, where the class member samples were treated as positive and the rest of the samples as negative. For each class specific learner we evaluated the so-called class accuracy. A test sample was classified to the class whose corresponding learner gave the highest score. The accuracy for the whole dataset was the ratio of the number of correctly classified samples and the total number of samples. The evaluation of the classification performance was carried out, among others, via standard receiver operator characteristic (ROC) analysis, which is based on the ranking of the objects to be classified [47]. This analysis is performed by plotting sensitivity versus *1*-specificity at various threshold values, and the resulting curve is integrated to give an area under the curve (AUC) value. For a perfect ranking AUC=1.0, while for a random ranking AUC=0.5.

2.4.3 Feature selection

A common goal in microarray data classification for diagnosis purposes is to select a minimal number of genes that could work as signatures for specific tumors. Since the SVM is generally thought to perform best in such classification problems, we introduce the *Recursive Feature Elimination* (RFE) algorithm, a

recently proposed feature selection method described in [48] which was designed in close relationship with SVM. The method seeks to recursively eliminate features, keeping the "best" m that lead to the largest margin of class separation using an SVM classifier. Considering the subset of surviving n features at a certain point in the procedure, the algorithm is basically the following:

1. Train the SVM with the n dimensional data, and thus obtain w.

2. Compute the feature ranking criteria $c_i=(w_i)^2$, $i=1,n$.

3. Find and eliminate the feature with the smallest ranking criterion $f=argmin(c)$.

The procedure is repeated until the number of remaining features reaches m. The RFE algorithm was used as part of the Spider package [67]. RFE was employed with a linear kernel SVM, included in the same software package.

2.4.4 Visualization

Visualization is an important topic in the analysis of high-dimensional measurements, especially because it facilitates the better understanding of the data. Here we shall only summarize three state-of-the-art graphical representation methods suitable for microarray data visualization.

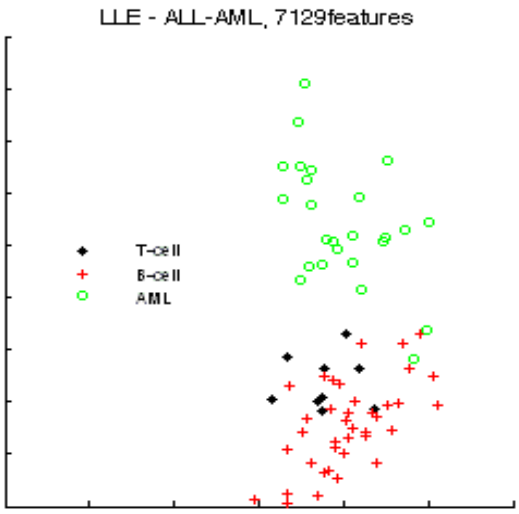

Figure 13. The LLE maping of high dimensional gene expression data into the 2D space.

The *Locally Linear Embedding* (LLE) is a distance preserving non-linear mapping from the high-dimensional original space into a lower dimensional space. Using this method [49] the dataset can be mapped into the 2D space, and thus easily plotted on a graphic as exemplified in Figure 13. The resulting two dimensions are abstract and do not correspond to any physical variable, therefore we omitted to annotate the axes. The colors correspond to classes. The method was used in Matlab, and the number of neighborhoods parameter required by the procedure was set to the number of samples. Another proposed visualization scheme is the *RadViz* [50] algorithm where the features (i.e. the genes) are represented as anchors that are equally spaced around the unit circle. The samples are then represented as points inside this unit circle. Their positions depend on the gene expression values: the higher the value for a gene, the more the anchor attracts the corresponding point. This method works best with relatively few (3–20) features, thus requiring a priori feature selection. Finally, the *Heat Map* with an optional hierarchical clustering on the genes can be also employed as a graphical

representation of the expression data matrices where the values taken by a feature are represented as color intensity in a 2D map. The visualizations were generally performed in Matlab, using the implementations provided by the authors of these methods.

2.5 Kalman Filtering – A Joint Perspective

This section presents the main contribution of this thesis. The procedure of molecular classification itself, as introduced earlier, is based on the fact that gene expression profiles work as surrogates for the biological state. Still, living cells are inherently dynamic; hence microarray measurements capture a large amount of expression variance. A large number of environmental error sources also corrupt the gene expression data, even though normalization procedures are meant to reduce such influences. These two types of variation alter the true gene expression states associated with the particular diseases in question. Under such circumstances the Kalman state estimator, embedded in a block diagram in Figure 14, provides a reasonable framework for preprocessing the expression data by removing the noise and estimating the multivariable noise-free tumor specific states.

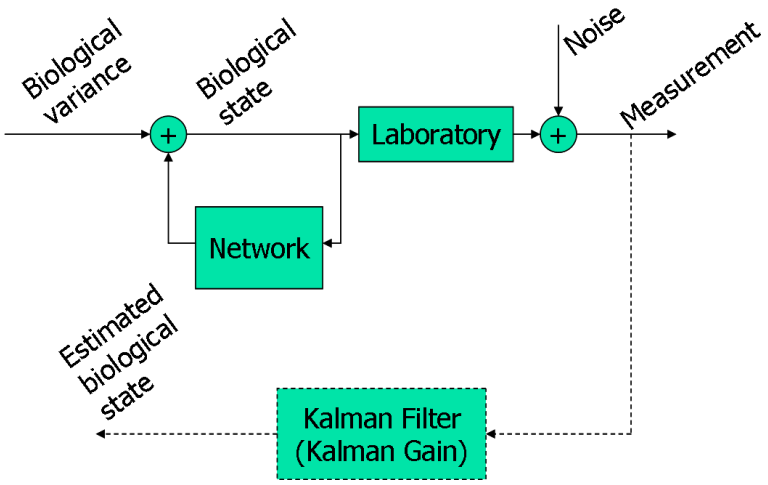

Figure 14. Block diagram of the biological state measurement with Kalman filtering.

The Kalman filter (KF) [51][52][53] is a powerful mathematical tool that has been widely used in many fields of engineering from systems and control theory to signal processing, due to its robustness even under the violation of the normality assumption. It has also been used in supervised learning as well as in myriads of real world applications. Its applications in the bioinformatics field however were limited [54], not taking advantage of its full potential as a multivariate signal processor. The KF is based on the assumption of a continuous system that can be modeled as a normally distributed random process X, with mean $\bar{x}$ (the state) and variance P (the error covariance):

$$X \sim N(\bar{x}, P) \tag{21}$$

The KF furthermore assumes that the output of the system can be modeled as a random process Z that is a linear function of the state $\hat{x}$ plus an independent, normally distributed, zero-mean white noise process V,

$$\bar{z} = H\bar{x} + \bar{v} \tag{22}$$

where, $V \sim N(0,R)$ and $E\{XV\}=0$. H represents the system output matrix. For our study we model the microarray data flow using the following simplified discrete time state-space representation of Equations (21) and (22):

$$\begin{aligned} x_k &= x_{k-1} + w_k \\ z_k &= x_k + v_k \end{aligned} \tag{23}$$

The first equation is a linear form of (21) containing the addition of an innovation process $W \sim N(0,Q)$. Vectors w_k and v_k may be interpreted as the modeling error (i.e. the deviation from a mean, stem-state towards the particular biological states in question) and measurement noise, respectively, the latter comprising the previously mentioned functional and experimental variances. Note that since the state transition matrix equals the unit matrix I, as does the output matrix H, they have been omitted for simplicity. The network block in Figure 14 corresponds to the state transition matrix. A discussion on how to integrate actual transcription network information is given in the Further Discussion section. Given the models of the white noise processes W and V (Q and R, respectively) and the array measurements z_k, the aim of the KF here is to estimate the state vectors $\hat{x}_k$ containing noise-free gene expression data. Considering the microarray profiling process as stationary (i.e. its statistical properties remain constant over time), the Kalman iterative estimation will converge to the steady-state KF, in which case the error covariance can be computed by solving the discrete algebraic Riccati equation (ARE):

$$P = P - P(P+R)^{-1}P + Q \tag{24}$$

Hence, the Kalman gain is given by:

$$K = (P+R)^{-1}P \tag{25}$$

The above equations are greatly simplified due to the omission of the state transition and output matrices for the same reason as noted previously. Finally, the estimated expression state vector is

$$\hat{x}_k = \hat{x}^- + K(Z_k - \hat{x}^-) \tag{26}$$

where, $\hat{x}^-$ is an estimate of $\bar{x}$ based on the previous samples. An important issue within Kalman filtering is the filter tuning. Given the training vector set, $\hat{x}^-$ can be chosen as the average of the class means, where for each class the means are computed from the member samples. We further use the training set to initialize and tune the two KF parameters, namely Q and R. To reduce the dimensionality of the problem, we propose the singular value decomposition [55]:

$$Z = UDY \tag{27}$$

The rows of Y are eigengenes and capture most of the variance of the original training dataset, while the columns correspond to the samples. The covariance matrix Q of the innovations can thus be obtained as the between-class covariance (i.e. the covariance of the class means with $\hat{x}^-$ subtracted) evaluated on the reduced dimensionality training set Y. The measurement noise model R is estimated as a weighted form of the within-class covariance of Y (i.e. the covariance of Y with the class means subtracted). To avoid over-fitting we tune these parameters by introducing some uncertainty variance such that $Q=Q+qI$ and $R= R+rI$. Our test runs led us to empirically conclude that in the case of single channel raw intensity array data (i.e. Affymetrix) $q=Q_{11}$ and $r=R_{11}$ are good choices for a reasonably good performance. Here the *11* index refers to the first eigengene usually considered as the offset of the microarray dataset, in which case it has a quite small variance. For expression log-ratio data (usually coming from

dual channel cDNA chips) or very sparse expression matrices these parameters yield acceptable results when we choose:

$$q = \sum_{i=1}^{n} Q_{ii}$$
$$r = \sum_{i=1}^{n} R_{ii} \tag{28}$$

n being the number of training samples. With the tuned parameters we compute the low dimensional Kalman gain K_Y using Equations (24) and (25). Finally, from Equations (26) and (27), the filtered gene-expression state vector is given by:

$$\hat{x}_k = \hat{x} + UDK_Y D^{-1} U^T (z_k - \hat{x}^-) \tag{29}$$

where, z_k now spans the entire dataset, including both train and test measurements. We implemented the actual expression data specific Kalman filter in Matlab and the source code is available on Kelemen et al.'s [57] supplementary information website.

Chapter 3

RESULTS AND DISCUSSION

3.1 Summary of the 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 are also summarized. 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 in Puskás et al [7], although these methods, or similar are used also in Nagy et al [56] and Zvara et al [8]. The results of Nagy et al [56] were used to present the custom application of the χ^2 test to assess for the effect of the amplification protocol used for sample preparation, on the detected expression changes. A more complex analysis of transcription profiles is exemplified using Zvara et al [8]. A hierarchical clustering was performed on a group a microarray data samples coming from both healthy and schizophrenic individuals. The unsupervised method discovered the two biological classes. At the same level of analysis complexity, in Kelemen et al [57] we are concerned with classification (supervised clustering). Here we apply the proposed Kalman filtering procedure on seven publicly available cancer expression datasets, and test several classification methods before and after filtering. A large, but mostly technical discussion of the Kalman filtering results with regard to classification is also provided and some other mathematical methods that were not introduced earlier are used here for this sole purpose.

3.2 Applied Bioinformatic Analyses for the Identification of the Genes Regulated by *NtrR* in *S. meliloti*

Here we aimed to identify the complete set of protein-coding genes influenced by loss of *ntr*R function in *Sinorhizobium meliloti* under aerobic and microaerobic conditions [7]. Microarray hybridizations were carried out to compare transcript levels in the wild type and mutant bacteria strains grown under both conditions. Mean signal and mean local background intensities were obtained for each of the 6207 spots on the arrays. Spots were flagged as "empty" if R was ≤ 1.5 in both channels, where *R*=(signal mean-background mean)/background standard deviation, and these were not included in the further analysis. A floor value of 20 was also used as threshold for the intensities. Data representing the log2 ratio of expression under microaerobic and aerobic conditions in both wild type and mutant strains were determined by cross-microarray comparisons, using single color intensities to calculate ratios. The duplicate experiments resulted in two average datasets calculated from triplicate spots representing each gene. Four combinations of ratios were calculated: wild type microaerobic/wild type aerobic; mutant microaerobic/ wild type microaerobic; mutant aerobic/wild type aerobic; mutant microaerobic/mutant aerobic. Before calculating the average ratios, tip-LOWESS normalization (i.e. LOWESS on each grid of the microarray) was performed for each case. Only those ratios were determined where both of the median intensities were above the 2SD of the background. Genes significantly up- or down-regulated were identified by *t*-statistics, using a significance threshold value of a=0.05. This work encompasses therefore the three basic steps required for the numerical analysis of a comparative microarray experiment: quality control, normalization and detection of regulated genes. The changes resulting from the microarray analysis were verified using QRT-PCR. As suggested by the results, the *ntr*R mutation affects genes encoding for various functions in symbiotic nitrogen fixation, transport, metabolism, or heat shock.

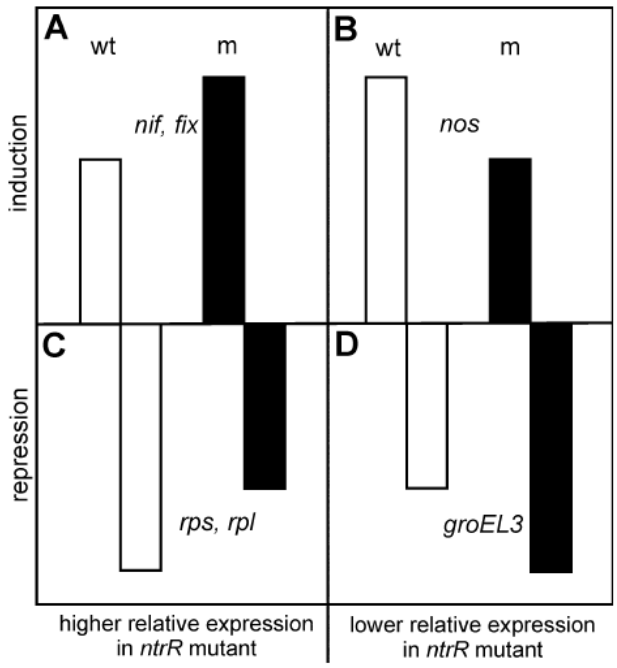

Figure 15. Schematic representation of the modulating effect of *ntr*R on transcription levels under microaerobiosis.

The cross-comparison reveals that some genes are induced under microaerobiosis, as shown in Figure 15, e.g. the members of the *nif*/*fix* cascade, which are up-regulated in the mutant relative to the wild type cells. The same figure shows that other genes, such as those participating in transcription-translation and biosynthetic processes (*rps*, *rpl*) were repressed primarily due to the condition (microaerobiosis), but were less affected in the mutant. Also, metabolic function encoding genes (*nos* family), were found to be induced by microaerobiosis, but somewhat repressed in the mutant. Some chaperonin genes like *groES3* were down-regulated under microoxic conditions, but in the mutant strain this effect was more pronounced than in the wild type cells.

3.3 Assessment of the Amplification Protocol Used in Sample Preparation on the Detected Gene Expression Changes

The objective of this study [56] was to infer the influence of the DNA amplification technique used, on the outcome of an expression measurement. A microarray experiment suite was carried out in order to identify the genes that express differentially between lipopolysacharide-treated and untreated mouse macrophages. Concerning the underlying nucleic acid sample amplification, two strategies were undertaken: an exponential phase amplification and a saturation phase over-amplification. A third protocol using dendrimer-based signal amplification was also employed for a control experiment. Out of the 3200 investigated genes, 15 were selected for QRT-PCR analysis and validation. The composition of this subset was balanced with regard to expression changes (i.e. it contained over-expressed, repressed, as well as un-regulated genes). Total RNA (1μg) was reverse transcribed and 15-ng aliquots were PCR amplified (Figure 16) with two protocols resulting in DNA samples from early phase (13th–15th cycles) and late exponential, early saturation phase (21st cycle). Thus, the two amplification strategies were again applied. Following the Pfaffl method, the QRT-PCR data was subjected to the one-sample *t* test to assess again the significance of the expression changes for the 15 selected genes. Within the control experiment that involved no DNA amplification, the same genes were selected and the data was subjected to the same treatment. Thus, to determine the effect of the amplification factor on the measured and categorized gene-expression changes, the χ^2 test, as described in the Methods section, was used. The overall significance threshold was set to 0.05. When comparing the exponentially amplified sample data to the control, a *p*-value of 0.8807 was obtained. No significant influence of the amplification could therefore be detected on the expression change composition. In the case of over-amplification, on the other hand, $p=0.0291$ suggests a strong distortion induced by the experimental factor on the expression data. This result is due to the fact

that, while the exponential phase sample amplification protocol preserves the original gene expression ratios that we want to reproduce at detectable levels, in the saturation phase these ratios tend to roughly drift toward 1, distorting the results.

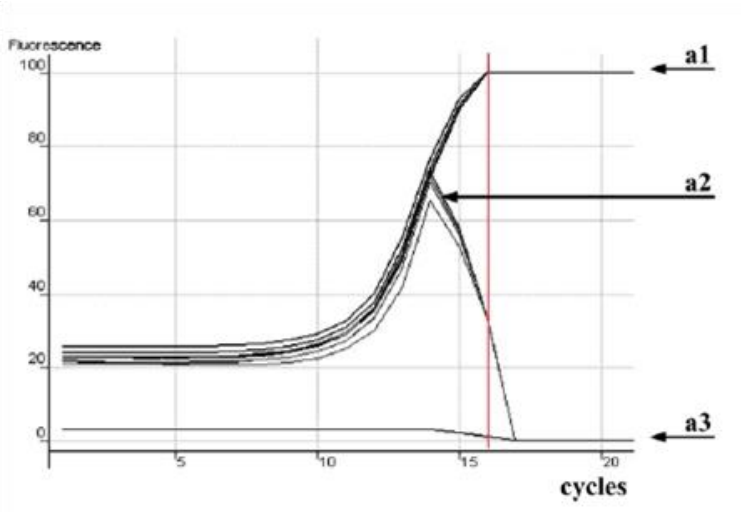

Figure 16. cDNA amplification with QRT-PCR of the LPS-treated mouse macrophage. With the QRT-PCR halted at the 14th cycle, the amplified cDNA (a2) was generated in the exponential phase of the reaction; the overamplified cDNA (a1) was isolated from reactions halted at the 21st cycle; a3 denotes the nontemplate control.

Clearly there are distortions that cannot be corrected by numerical means, such as those that appear here, during sample preparation. This type of noise, which physically influences the biological sample, has to be corrected on the protocol level. As a result of this statistical analysis, the exponential phase amplification was proposed as a better alternative for reliability and increased reproducibility.

3.4 Schizophrenia Diagnosis and Marker Genes

13 drug-naive schizophrenic patients and 10 control individuals were screened to identify novel peripheral genetic markers of schizophrenia [8]. A cDNA microarray analysis was performed in order to pre-screen for expression

regulation patterns on peripheral blood lymphocytes, and to identify potential peripheral marker genes. Out of the 3200 clones that were present on-chip, two were selected, based on their differential expression. These genes, namely *DRD2* and *Kir2.3*, also showed strong correlation with the disease. A validation experiment has been performed by means of QRT-PCR on these two features. We finally performed a hierarchical agglomerative clustering on the obtained two-dimensional data (based on the two mentioned genes only). As pictured in Figure 17, the procedure clearly delineates the schizophrenia from the normal healthy samples based on the proposed two-gene signature.

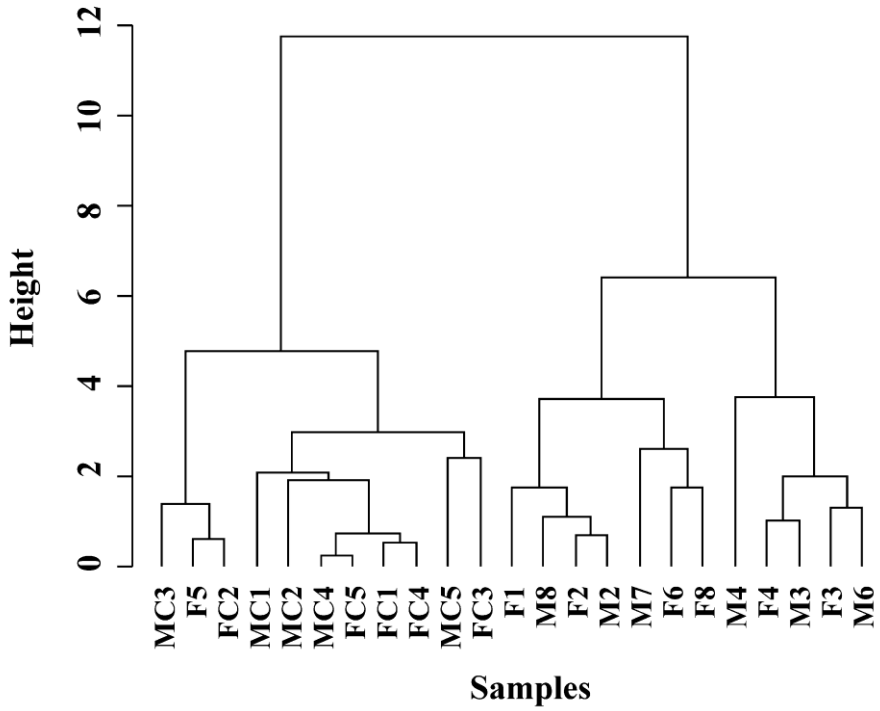

Figure 17. The hierarchical clustering based on the reduced expression dataset clearly separates the two main clusters (MC-male control, FC-female control, M-male patient, F-female patient).

A biological interpretation of the identified schizophrenia signature is given in the following. The increased occupancy of the D2 subclass of dopamine receptors by

dopamine is one of the hypotheses explaining the nature of schizophrenia. DRD2 belongs to this class of receptors and is coupled to a G-protein. Receptor-activated G-proteins can either activate or inactivate inwardly rectifying potassium channels, such as Kir2.3. Several different potassium channels are involved in electrical signaling in the nervous system. The malfunctioning of the K^+ channels has also been brought in association with schizophrenia.

3.5 Kalman Filtering for Disease-State Estimation

We propose using the Kalman filter (KF) as a pre-processing step in microarray-based molecular diagnosis [57]. Here, we show that employing the KF 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. We demonstrate thus the utility and performance of the KF as a robust disease-state estimator on publicly available binary and multiclass microarray datasets in combination with the most widely used classification methods to date. Moreover, using popular graphical representation schemes we show that our filtered datasets also have an improved visualization capability.

3.5.1 Datasets

We tested the Kalman filtering-classification scheme on a number of publicly available datasets, which are summarized in Table 2. The leukemia (ALL-AML) dataset of [29] is a popular dataset and is often used to test binary classification algorithms. Using the original sample annotation we partitioned this dataset into three leukemia classes. Hence the dataset consisted of T lineage acute lymphoblastic leukemia (T-ALL), B lineage acute lymphoblastic leukemia (B-ALL) and acute myeloid leukemia (AML) samples.

Name	#classes	#genes	#train	#test	Source
ALL-AML	3	7129	38	34	Golub et al. 1999
Tumors	14	16063	144	54	Ramaswamy et al. 2001
MLL	3	12582	57	15	Armstrong et al. 2002
LC	2	12533	32	149	Gordon et al. 2002
SRBCT[a]	4	2308	63	25	Khan et al. 2001
BC[b]	2	24481	78	19	van't Veer et al. 2002
Leukemia[1]	7	12558	215	112	Yeoh et al. 2002

Table 2. Features of the datasets.

In our study we included two other leukemia datasets: the mixed lineage leukemia (MLL) dataset [58] and the pediatric acute lymphoblastic leukemia (Leukemia) dataset [61]. The former consists of acute lymphoblastic leukemia (ALL) and AML samples along with ALLs carrying a chromosomal translocation involving the MLL gene. The latter is composed of B-ALL subtypes expressing BCR-ABL, E2A-PBX1 and TEL-AML1, respectively, a hyper-diploid karyotype, as well as MLL, T-ALL and a novel leukemia subtype. The "various tumor types" (Tumors) dataset [27] is considered a difficult dataset and consists of 14 classes of tumors: breast, prostate, lung, colorectal, lymphoma, bladder, melanoma, uterus, leukemia, renal, pancreas, ovary, mesothelioma and central nervous system tumors. The dataset (LC) of [59] contains microarray data that accounts for two distinct pathological alterations of the lung: malignant pleural mesothelioma and adenocarcinoma. The small, round blue cell tumors (SRBCT) of childhood dataset [40] includes a training set of neuroblastoma, rhabdomyosarcoma, Burkitt lymphoma and the Ewing family of tumors samples and an independent test set that, besides the samples belonging to the training classes, also contains samples that should not be classified into any of these tumor types. [60] provides a dataset (BC) consisting of samples coming from breast cancer patients that were clustered by the original authors into two classes according to the patient's response to adjuvant therapy: relapse and non-relapse.

[a] dataset containing log-ratio expression data

[b] sparse dataset

3.5.2 Classification results

We applied the Kalman filtering on the described datasets and for a comparative study SVM, ANN, 1NN and RF supervised learning methods were evaluated in full gene set manner. Table 3 summarizes the Accuracy and ROC scores we obtained. Evidently, the KF definitely improves the classification results of the ANN, 1NN and RF. The SVM results were boosted in 64% of the overall scores. We should mention that, in the four-class SRBCT dataset there were 25 test samples, but among the test elements there were 5 samples which were not members of any of the training classes. We expected each of the class specific learners to reject these samples. The procedure however, will necessarily assign them to the closest classes, which results in an apparent decrease of performance. Owing to these 5 cases, for the SRBCT dataset the mean of the class accuracies was shown.

		SVM			ANN		1NN		RF	
		Original	PCA	KF	Original	KF	Original	KF	Original	KF
ALL-AML	ROC score	**0.99**	**0.99**	**0.99**	0.97	**0.99**	0.73	**1**	0.92	**0.95**
	Accuracy	0.91	0.82	**0.97**	0.91	**1**	0.82	**1**	0.74	**0.94**
BC	ROC score	**0.88**	0.81	0.70	0.67	**0.74**	0.23	**0.68**	0.64	**0.68**
	Accuracy	0.58	0.63	**0.68**	0.37	**0.74**	0.63	**0.63**	0.63	0.63
Leukemia	ROC score	0.97	0.96	**0.98**	0.90	**0.98**	0.60	**0.88**	0.94	**0.96**
	Accuracy	0.50	0.29	**0.7**	0.37	**0.58**	**0.89**	0.87	**0.86**	0.76
LC	ROC score	**1**	0.99	0.99	**1**	0.99	0.59	**0.99**	**0.99**	**0.99**
	Accuracy	**0.99**	0.98	0.98	**0.99**	0.98	0.94	**0.98**	0.93	**0.98**
MLL	ROC score	**1**	**1**	**1**	**1**	**1**	0.87	**1**	0.92	**0.98**
	Accuracy	**1**	**1**	**1**	**1**	**1**	0.93	**1**	0.8	**1**
SRBCT	ROC score	0.99	0.99	**1**	0.99	**1**	0.66	**1**	0.99	**1**
	Accuracy[c]	0.97	0.97	**0.99**	0.94	**0.95**	0.91	**0.95**	0.93	**0.98**
Tumors	ROC score	**0.95**	0.91	0.94	0.90	**0.94**	0.72	**0.92**	0.84	**0.87**
	Accuracy	0.74	0.63	**0.80**	0.50	**0.80**	0.46	**0.67**	0.48	**0.67**

Table 3. Comparison of the classification performance on the original and the Kalman filtered datasets. The best performing value for each method is shown in bold, and the overall best values are also underlined.

[c] denotes the mean of the class accuracies

To assess the significance of filtering on microarray data classification we performed paired two sample *t*-tests to compare the accuracies and ROC scores of the classification procedures on the original datasets with their counterparts in the KF case. The *t*-statistic was applied in one-tail fashion testing against the alternative hypothesis that the mean of accuracies/ROC scores produced by a certain method on the raw datasets is less than the mean of the matched performance measures on the pre-processed datasets. Table 4 shows that with 95% confidence the KF approach significantly improves the accuracy or the ROC score. In our study we also compared the KF scheme with a different approach to multivariate filtering. The principal component analysis (PCA) based filtering consists of removing the non-significant variance components computed using the eigen-decomposition of the covariance matrix of the training set.

t-Test (α=0.05)	Accuracies	ROC scores
$p_{SVM \geq KF+SVM}$	0.033	0.18
$p_{PCA+SVM \geq KF+SVM}$	0.043	0.35
$p_{ANN \geq KF+ANN}$	0.028	0.03
$p_{1NN \geq KF+1NN}$	0.033	0.0002
$p_{RF \geq KF+RF}$	0.052	0.005
$p_{SVM \geq PCA+SVM}$	0.083	0.058

Table 4. Significance test results

The PCA results with SVM are shown in Table 3. As opposed to PCA the KF retains the dataset in the original gene space and is also supervised procedure from a classification point of view. This point is made clear by the *p*-values in Table 4. In the SVM framework, the PCA filtered datasets did not yield any improvement at a significance level of 0.05 in accuracy/ROC score compared to the original data. Using the same learning algorithm, the KF shows significant accuracy increase over the PCA technique. The advantage of such a pre-processing approach here is not just a better classification performance, but also an improved visualization capability of the data.

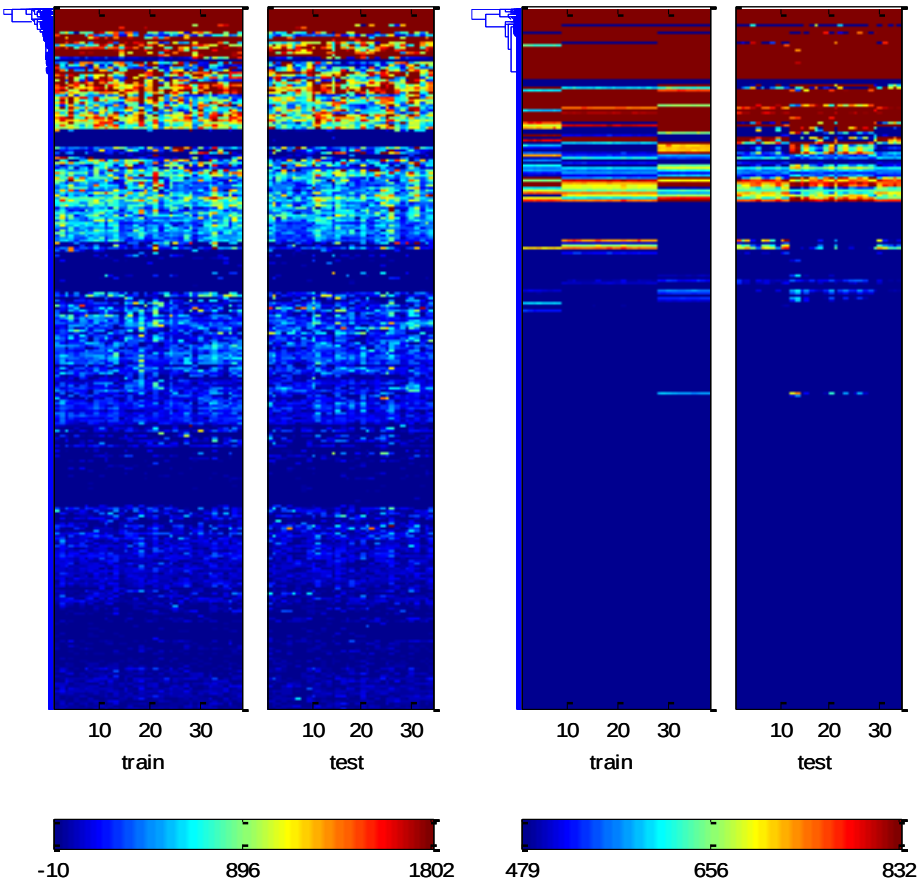

Figure 18. The heat map representation of the AML-ALL dataset. The first pair shows the original dataset and the second pair shows the filtered dataset.

The heat map with a hierarchical clustering presented in Figure 18 demonstrates how effectively the KF technique performs. The columns represent the samples and the clustering was effectuated on the genes (the rows). Each gene expression value is encoded by a color according to the legend below the heat-map. A visual inspection on the original dataset on the left shows no distinction of the classes due to noise. Filtering helps remove noise and the leukemia classes become visible. The standard deviation of the gene expression values was reduced in each

class. And the genes that carried no information related to the class separation were homogenized.

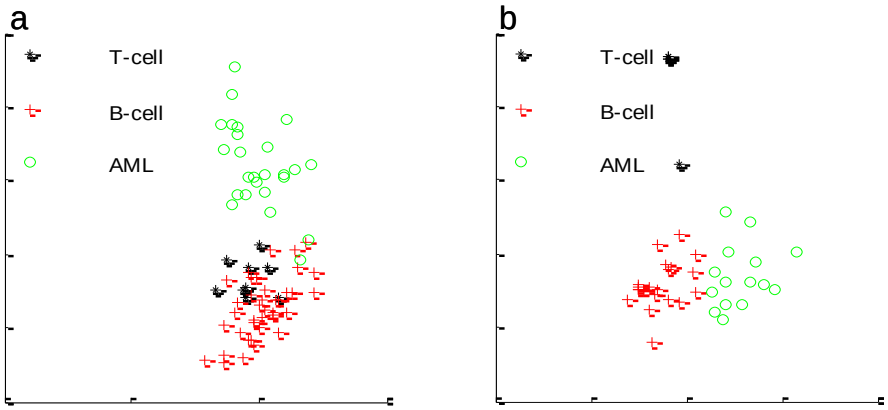

Figure 19. The original (a) and the Kalman filtered (b) AML-ALL dataset visualized by LLE.

Another type of visualization underlines the same performance of the Kalman filter. Figure 19a depicts the original AML-ALL dataset while Figure 19b depicts the Kalman filtered dataset. As we mentioned in the Methods section, the axes here stand for two abstract dimensions which result from the locally linear embedding. These two dimensions, obtained from the reduction of the 7129 genes, do not correspond to any physical quantity or variable, and therefore are not named on the figure. The classes within the 2D points are marked distinctly. The LLE representation clearly shows that the classes are more delineated with filtering than without.

3.5.3 Signature features

The RFE feature selection method was evaluated on the original and the Kalman filtered datasets to test whether filtering could help find more reliable subsets of tumor marker genes.

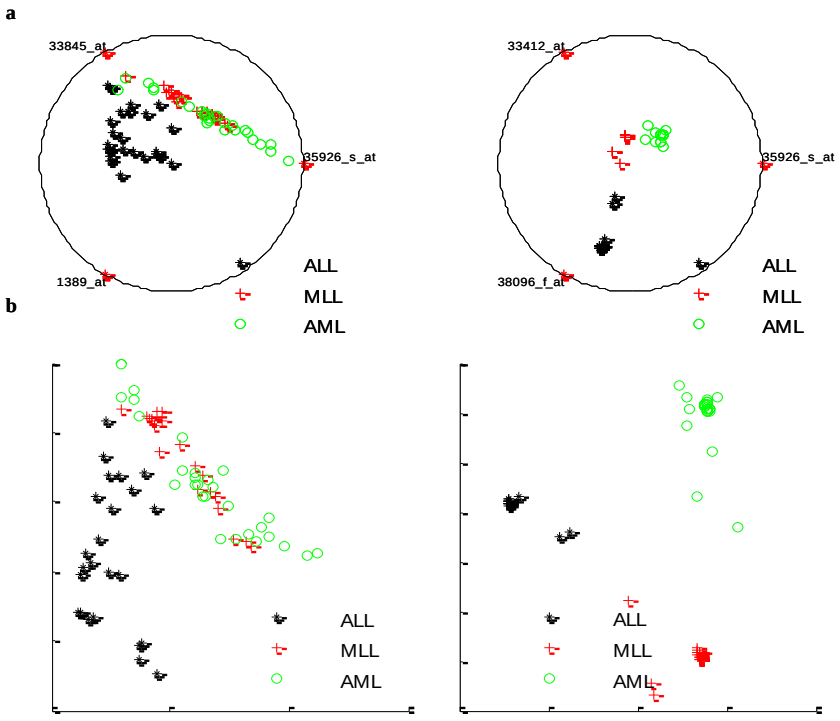

Figure 20. Visualization of the original (left side) and the Kalman filtered (right side) MLL dataset. In (a) the RadViz method was used on three genes selected by RFE and plotted on the unit circle. The same genes were used with LLE in (b).

The results we obtained, summarized in Table 5, 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. The same result is noticeable in Figure 20 where, in a three-best-feature setup, the MLL classes are well separated in the KF data but they are overlapped in the original vector set. Figure 21 shows a heat map visualization of the MLL dataset with 50 selected features.

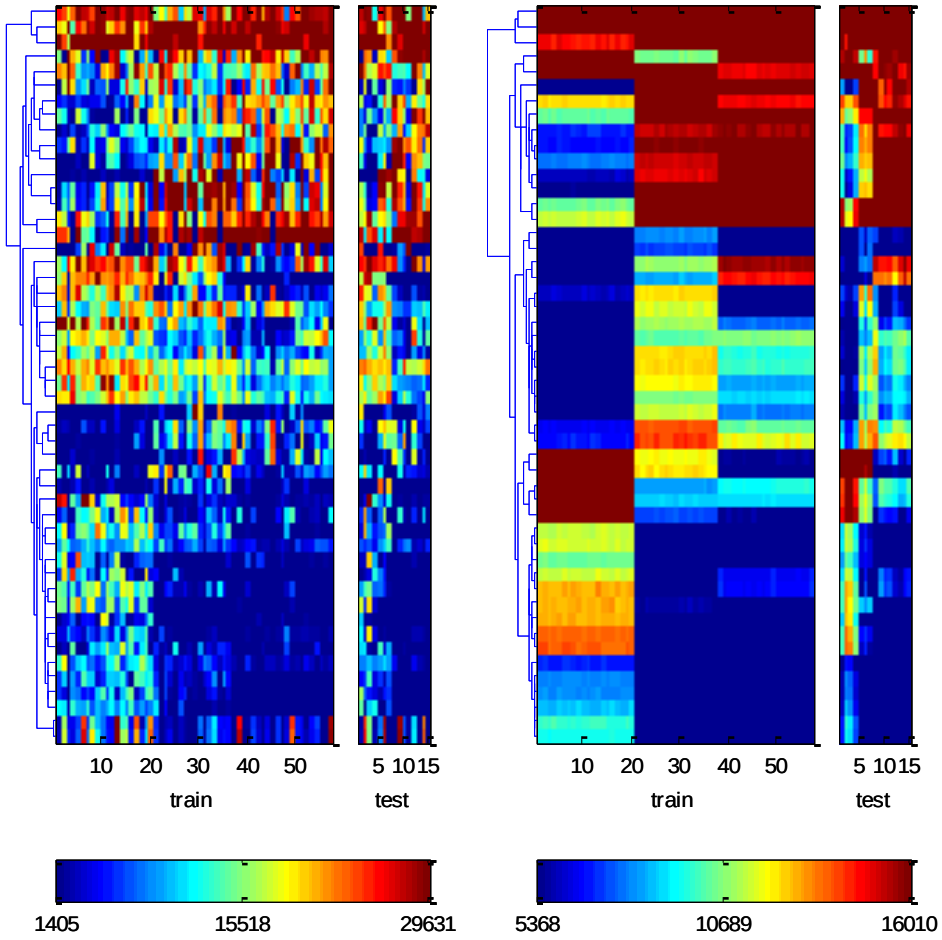

Figure 21. Heat map of the best 50 genes selected by RFE from the MLL dataset. On the Kalman filtered dataset (right) the features are less noisy and the three classes are further apart than in the original dataset (left).

These genes were selected so that their expression is in close (numerical) relationship with the leukemia subtypes. The classes are almost visible now even on the raw data. While on the train set KF obviously removes the measurement noise, which results in clearly separated tumor groups, the variance of the test set is also noticeably diminished by the filter. Note that the selected genes from the original and the filtered datasets are distinct.

Score	Dataset Name		Number of selected features with RFE								
			2	3	5	7	10	15	20	30	50
Accuracy	ALL-AML	Original	0.53	0.56	0.68	0.68	0.68	0.65	0.74	0.76	0.85
		KF	0.74	0.94	0.82	0.85	0.97	1	0.97	0.97	0.97
	BC	Original	0.79	0.63	0.63	0.63	0.63	0.63	0.58	0.58	0.58
		KF	0.63	0.63	0.63	0.63	0.63	0.58	0.58	0.58	0.63
	Leukemia	Original	0.26	0.46	0.58	0.60	0.66	0.82	0.75	0.78	0.77
		KF	0.19	0.32	0.59	0.68	0.79	0.79	0.77	0.81	0.54
	LC	Original	0.95	0.98	0.99	0.99	0.97	0.98	0.97	0.97	0.98
		KF	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98
	MLL	Original	0.67	0.67	0.73	0.87	0.87	0.87	0.93	1	0.93
		KF	1	1	1	1	1	1	1	1	1
	SRBCT	Original	0.81	0.74	0.84	0.81	0.81	0.85	0.89	0.92	0.97
		KF	0.88	0.88	0.95	0.99	0.99	0.99	0.99	0.99	0.99
	Tumors	Original	0.13	0.11	0.19	0.24	0.26	0.43	0.50	0.46	0.54
		KF	0.17	0.17	0.35	0.48	0.52	0.65	0.65	0.69	0.74
ROC	ALL-AML	Original	0.68	0.65	0.83	0.83	0.89	0.87	0.90	0.92	0.95
		KF	0.87	0.93	0.92	0.94	0.99	0.99	0.99	0.99	0.99
	BC	Original	0.89	0.81	0.76	0.79	0.75	0.73	0.75	0.62	0.78
		KF	0.69	0.69	0.68	0.68	0.68	0.68	0.68	0.68	0.68
	Leukemia	Original	0.74	0.82	0.88	0.89	0.90	0.95	0.93	0.92	0.95
		KF	0.74	0.84	0.95	0.96	0.98	0.98	0.99	0.98	0.98
	LC	Original	0.97	0.99	1	1	1	1	1	1	1
		KF	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
	MLL	Original	0.86	0.88	0.93	0.98	0.99	0.96	0.95	1	1
		KF	1	1	1	1	1	1	1	1	1
	SRBCT	Original	0.84	0.79	0.90	0.90	0.89	0.93	0.98	0.97	0.99
		KF	0.92	0.93	0.97	1	1	1	1	1	1
	Tumors	Original	0.61	0.65	0.75	0.79	0.80	0.81	0.85	0.84	0.88
		KF	0.68	0.77	0.86	0.89	0.87	0.90	0.91	0.91	0.93

Table 5. The accuracies and ROC scores obtained via SVM depending on the number of selected features.

The names of the two best performing genes within the filtered MLL dataset are given in Table 6.

Clone ID	Accession	Description
33412_at	AI535946	vicpro2.D07.r conorm Homo sapiens cDNA 5', mRNA sequence
38096_f_at	M83664	Human MHC class II lymphocyte antigen (HLA-DP) beta chain mRNA

Table 6. The two best performing MLL markers.

The linear or nonlinear combination of these genes' expression levels does not necessarily mean an actual relationship between them. In fact, the KF uses the variance of all the involved features in estimating the expression state, and these genes may just be the "top of the stack" or the finely regulated distant ends of the network. The database contains little information on vicpro2, although it has come up as marker gene candidate in many classification projects in the literature. It was associated with prostate tumor. The major histocompatibility complex genes (lymphocyte antigen) are involved in the immune response. So there is a double association of the selected genes with tumor in general and leukemia in particular. To compare the quality of features selected from the original datasets with the filtered ones, the fisher separation ratio (*FSR*) was used. The *FSR* is a scalar which is large when the between-class covariance is large and when the within-class covariance is small. Out of the many possible choices of criterion [35] our ratio was defined as $FSR = Tr\{S_W^{-1} S_B\}$, where $Tr\{\}$ denotes the trace of a matrix and S_B and S_W are the between- and within-class scatter matrices, respectively [62]. Here the between-class scatter matrix is the scatter of the class mean vectors around the overall mean vector, while the within-class scatter matrix denotes the weighted average scatter of the covariance matrices of the sample vectors belonging to each class. Table 7 lists the FSR scores for 10 features independently selected from each dataset.

Dataset	Original	KF
ALL-AML	14.088	19.737
BC	1.480	2.677
Leukemia	4.079	66.299
LC	5.757	4.164
MLL	8.481	67.659
SRBCT	3.621	105.181
Tumors	3.406	29.668

Table 7. FSR on 10 features selected via RFE ($p_{\text{Original}\geq\text{KF}} = 0.0245$).

The significantly larger scores (p=0.0245 obtained from a t-test, as described previously) produced by the KF features demonstrate the greater predictive power of the estimated expression data that best define the causal biological states. In conclusion, the KF is a systemic approach to filtering, each gene's expression being estimated using the variances of all the individual features, of course assuming that many genes reflect the biological state of the sample due to the transcriptional network. Hence, it remains for further study (i.e. PCR analysis) to assess whether the selected features can also independently predict and diagnose a tumor outcome.

Chapter 4

FURTHER DISCUSSION AND CONCLUSIONS

4.1 Further Study Perspectives – Beyond the Single Dataset

Since we are concerned with biological and typically gene expression data analysis methodology, in the following we shall discuss the possibility of extending KF procedure in the systems biological sense. The preliminary results given here are presented solely for the purpose of practical exemplification of the thoughts and ideas discussed. As we saw in the Methods section, so far the filter has been used in conjuncture with the most simplistic model of the microarray process. The model was driven stochastically only by random processes. It was also clear, that the filter is able to process transcriptome-wide data. Therefore, the question that arises here is how can we integrate information about the true transcriptional network, in its entirety, into this model and the filtering procedure itself? It is only natural for this sort of problems to emerge as we approach system level analysis, so particular to systems biology. Further, system level understanding of cancerous cells could provide deterministic and reliable strategies of effective treatment. One possibility of model enrichment is to expand the state-space equations in (23) to the more general form:

$$\begin{aligned} x_k &= Ax_{k-1} + Bu_k + w_k \\ z_k &= x_k + v_k \end{aligned} . \tag{30}$$

A is the state transition matrix, while B is the control matrix. The state transition matrix should account for the networked relationships between the states (i.e. transcripts), as well as the network dynamics. The acquiring of the matrix A can be done by estimation from time-series data. For this purpose, in our preliminary analysis we used a time-series dataset from Whitfield et al. [63]. This dataset was

obtained by expression profiling of a *HeLa* cell culture synchronized by arrest in S phase using a double thymidine block. As target tumor data the SRBCT dataset was chosen. Previous to any analysis, the two datasets were synchronized to each other, such that only common genes were kept. Also, the missing values within the Whitfield dataset were estimated using a *k*NN based algorithm described in [64]. Out of the remaining number of genes, we selected the 30 best ranked ones, based on the RFE-SVM results in Section 3.2.3. Based on the rough presumption that the SRBCT cell can be obtained from a *HeLa* cell by controlling the expression-states, we can write the following equations involving actual gene-expression data:

$$\begin{aligned} h_i &= Ah_{i-1} + B \cdot 0 \\ s_k^{\mu} &= As_k^{\mu} + B \cdot k \end{aligned}, \tag{31}$$

Here, h_i stands for the *i*-th *HeLa* transient response sample, while s_k^{μ} represents the average of the SRBCT samples belonging to the *k*-th (*k=1:n*) class (a steady state response). This black-box system identification problem can be solved by the least-squares procedure [1], when the number of samples is satisfactory:

$$(AB)^T = \left[\begin{pmatrix} h_i s_k^{\mu} \\ 0k \end{pmatrix} \begin{pmatrix} h_i s_k^{\mu} \\ 0k \end{pmatrix}^T \right]^{-1} \begin{pmatrix} h_i s_k^{\mu} \\ 0k \end{pmatrix} \left(h_i s_k^{\mu}\right)^T . \tag{32}$$

In our case the number of samples was sufficient for the system identification. Having obtained the system model (*A*, *B*), the Kalman filtering can be performed similarly as described in Section 2.5, except the ARE becomes:

$$P = A[P - P(P+R)^{-1}P]A^T + Q \tag{33}.$$

Also, the uncertainty parameters q and r may be dropped. Figure 22 shows the dataset filtered using the proposed procedure. In this preliminary study, the test samples known to be members of neither of the training classes were removed prior to classification. Thus, the classification using SVM was 100% accurate.

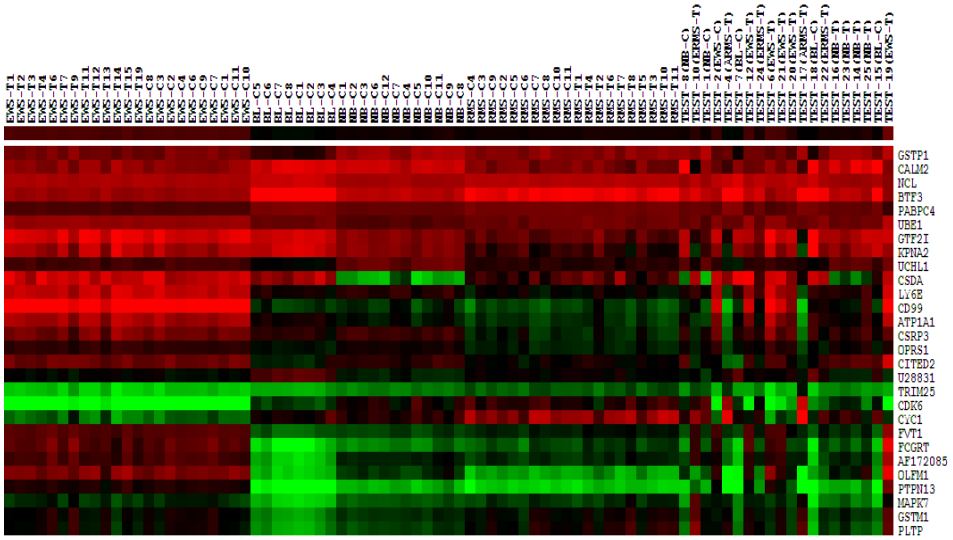

Figure 22. The filtered SRBCT dataset.

This promising result, which could not be obtained based solely on the raw SRBCT dataset, suggests that the transcription-network model of a few genes can roughly account for the entire system under certain circumstances. Figure 23, on the other hand, shows that a cluster analysis on the filtered full dataset can delineate the classes, and the "foreign" samples are quite differentiated as well. Such a system level analysis has several implications in classification. Some of the variance of the tumor samples may be identified as being of biological origin. Thus, the method can handle such variances and this is reflected on the classification results as well. More importantly, it is expected that the selected marker genes would also be more reliable. In the future, issues like system controllability could be also inferred, possibly leading to the identification of

actual drug target genes, which control the most of the cell events, and optimal control based treatment strategies could be employed.

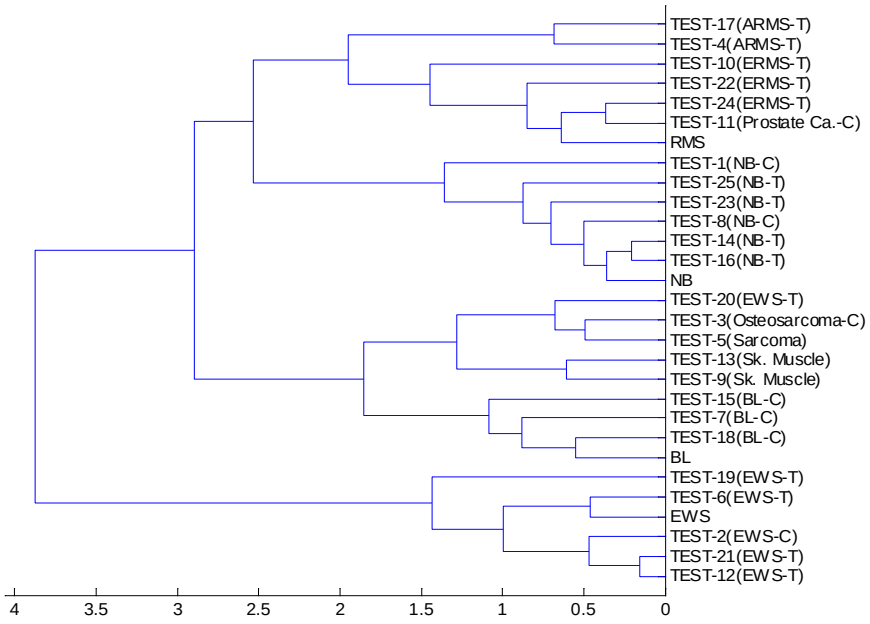

Figure 23. Cluster analysis on the filtered SRBCT dataset. EWS, BL, NB and RMS stand for the class means.

4.2 Conclusions

Data preprocessing is compulsory before biological interpretation of QRT-PCR and DNA microarray data. Some of our biologically interpreted result could also be verified in the literature showing that the preprocessing was effective and that the novel biological results are reliable. There are error that cannot be corrected numerically, such as those induced by sample preparation. The less distorting protocol should be used for these work phases. As we saw, the Kalman filter is a

powerful data processing tool. In the frame classification it performed well improving the accuracy of the used machine learning algorithms, and thus increasing the reliability of cancer diagnosis. The different levels of performance improvement on the different classification methods result from the nature, mathematical background and complexity of these methods. This is reflected in how well they can handle noise themselves. For example, the nearest neighbor algorithm is one of the simplest classifiers. It performs therefore quite poorly in noisy environment. On the linearly classifiable datasets, yielded by filtering, its performance is significantly improved. The different datasets were produced by different laboratories, probably using different protocols as well. In addition to that, there are various array platforms that the data come from. All these influence the noise estimates used by the Kalman filtering procedure. The KF procedure works best with normally distributed noise, although being quite robust to other distributions up to a certain degree. This clearly influences the performance on certain datasets. The results obtained within the classification frame intuitively lead to the idea of the KF being used also for the purpose of general expression-data normalization, in the broader sense. The only problem consists of estimating the measurement-experimental noise. This could be achieved for example by performing multiple technical repeats, prior to the actual experiment. Once passed over this obstacle, the procedure can in theory filter systematic as well as random noise, and thus could replace several steps of the by now conventional microarray data analysis. The technique is suitable for both QRT-PCR and microarray data, since these data are of the same nature. Actually, for the QRT-PCR data the implementation of the filter could simpler, since the number of investigated genes is smaller, thus the dimensionality of the problem is lower. The performance of the KF technique depends essentially on the tuning of the covariance matrices Q and R. In our implementation we used a flexible parametric setting, which allows us to handle the uncertainty of the noise estimates (due to the high dimensionality, the test samples' noise may be marginal

in the noise distribution). We tried the make these settings as general as possible, and yet provide overall good performance. Our choice of parameters proved to be reasonable for classification, although an improvement based on larger training data or better tuning formulae is possible. The filtering of one dataset took only a few seconds of CPU time, hence the technique is a fast and scalable method for pre-processing the gene-expression data.

REFERENCES

[1] Eykhoff P. (1974) System Identification. J. Wiley, London.

[2] Hartwell L.H. et al. (2003) Genetics: From Genes to Genomes, Second Edition. McGraw−Hill, New York.

[3] Lodish H. et al. (2004) Molecular Cell Biology, 5th edn. W. H. Freeman and Company, New York.

[4] Higuchi R. et al. (1993) Kinetic PCR: Real time monitoring of DNA amplification reactions. *Biotechnology*. **11**: 1026–1030.

[5] Schena M. et al. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*. **270** (5235): 467-70.

[6] Brazma A. et al. (2001) Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nature Genetics*, **29**: 365-371.

[7] Puskas L.G. et al. (2004) Wide-range transcriptional modulating effect of ntrR under microaerobiosis in Sinorhizobium meliloti. *Mol Genet Genomics*. **272**(3): 275-89.

[8] Zvara A. et al. (2005) 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*. **21**(2): 61-9.

[9] Darvas F. et al. (2004) Recent advances in chemical genomics. *Curr Med Chem*. **11**(23): 3119-45.

[10] Vass L. et al. (2006) Medium-throughput microarray-based approach for toxicogenomic profiling of small molecules. *QSAR & Combinatorial Science*. **25**(11): 1039-1046.

[11] Kitano H. (2001) Foundations of Systems Biology. MIT Press, Cambridge, Massachusetts.

[12] Berrar D.P. et al. (2003) A practical approach to microarray data analysis. Kluwer Academic Publishers, Norwell, MA.

[13] Knudsen S. (2002) A biologist's guide to analysis of DNA microarray data. John Wiley & Sons, New York.

[14] Tseng G.C. et al. (2001) Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Res*. **29**: 2549-57.

[15] Hoyle D.C. et al. (2002) Making sense of microarray data distributions. *Bioinformatics*. **18**: 576-84.

[16] Pfaffl M.W. (2001) A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Res*. 29: e45.

[17] Badiee A. et al. (2003) Evaluation of five different cDNA labeling methods for microarrays using spike controls. *BMC Biotechnol*. **11**: 3-23.

[18] Heller R.A. et al. (1997) Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. *Proc Natl Acad Sci USA*. **94**: 2150-2155.

[19] Altman N. (2005) Replication, variation and normalisation in microarray experiments. *Appl Bioinformatics*. **4**(1): 33-44.

[20] Chen Y. et al. (1997) Ratio-based decisions and the quantitative analysis of cDNA microarray images. *J. Biomed. Optics*. **2**: 364-374.

[21] Yang Y.H. et al. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research*. 30: el5.

[22] Cleveland W.S. and Devlin S.J. (1988). Locally weighted regression: an approach to regression analysis by local fitting. *J. Am. Stat. Assoc*. **83**: 596-610.

[23] Dudoit S. et al. (2002) Statistical methods for identifying genes with differential expression in replicated cDNA microarray experiments. *Statistica Sinica*. **12**: 111-139.

[24] Petrie A. and Sabin C. (2000) Medical statistics at a glance. Blackwell Science Ltd. , London.

[25] Patefield W.M. (1981) Algorithm AS159. An efficient method of generating r x c tables with given row and column totals. *Appl. Stat*. **30**: 91–97.

[26] Eisen M.B. et al. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*. **95**: 14863-14868.

[27] Ramaswamy S. et al. (2001) Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl. Acad. Sci. USA*. **98**, 15149–15154.

[28] Welcsh P.L. et al. (2002) BRCA1 transcriptionally regulates genes involved in breast tumorigenesis. *Proc. Nat. Acad. Sc. USA*. **99**(11): 7560-5.

[29] Golub T.R. et al. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*. **286**: 531–537.

[30] Vapnik V.N. (1998) Statistical Learning Theory. John Wiley & Sons, NY.

[31] Jaakkola T. et al. (1999) Using the Fisher kernel method to detect remote protein homologies. *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, California, pp. 149–158.

[32] Brown M.P.S. et al. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl Acad. Sci. USA*. **97**: 262–267.

[33] Zien A. et al. (2000) Engineering support vector machine kernels that recognize translation initiation sites. *Bioinformatics*. **16**: 799–807.

[34] Noble W.S. (2004) Support vector machine applications in computational biology. In Schoelkopf,B., Tsuda,K. and Vert,J.-P. (eds), *Kernel methods in computational biology*. Cambridge, MA, MIT Press, pp. 71–92.

[35] Bishop C.M. (1995) Neural Networks for Pattern Recognition. Clarendon Press, Oxford.

[36] Hertz J. et al. (1991) Introduction to the Theory of Neural Computation. Addison-Wesley, Redwood City, CA.

[37] Rumelhart D.E. et al. (1986) Learning internal representations by error propagation. In Rumelhart,D.E. and McClelland,J.L. and the PDP Research Group (eds), *Parallel Distributed Processing: Explorations In The Microstructure Of Cognition*, Volume 1: Foundations. MIT Press, Cambridge, MA, pp. 318–362.

[38] Baldi P. and Brunak S. (2001) Bioinformatics: The Machine Learning Approach. MIT Press, Cambridge, MA.

[39] Murvai J. et al. (2001) Prediction of protein functional domains from sequences using artificial neural networks. *Genome Res*. **11**: 1410–1417.

[40] Khan J. et al. (2001) Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat. Med.* **7**: 673–679.

[41] Dudoit S. et al. (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc.* **97**: 77–87.

[42] Fix E. and Hodges J. (1951) Discriminatory analysis, nonparametric discrimination: Consistency properties. Technical report, USAF School of Aviation Medicine, Randolph Field, Texas.

[43] Liao L. and Noble W.S. (2003) Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships. *J. Comput. Biol.* **10**: 857–868.

[44] Remlinger S.K. (2003) Introduction and application of random forest on high throughput screening data from drug discovery. http://www.samsi.info/200304/dmml/kickoffpresentations/remlinger.pdf

[45] Shi T. et al. (2005) Tumor classification by tissue microarray profiling: random forest clustering applied to renal cell carcinoma. *Mod. Pathology*. **18**: 547–557.

[46] Breiman L. (2001) Random forests. *Mach. Learn*. **45**: 5–32.

[47] Gribskov M. and Robinson N.L. (1996) Use of Receiver Operating Characteristic (ROC) analysis to evaluate sequence matching. *Comput. Chem.* **20**: 25–33.

[48] Guyon I. et al. (2002) Gene selection for cancer classification using support vector machines. *Mach. Learn*. **46**: 389–422.

[49] Roweis S. and Saul L. (2000) Nonlinear dimensionality reduction by locally linear embedding. *Science*. **290**: 2323–2326.

[50] Brunsdon C. et al. (1998) An investigation of methods for visualizing highly multivariate datasets. *Case Studies of Visualization in The Social Sciences*, pp. 55–80.

[51] Kalman R.E. (1960) A new approach to linear filtering and prediction problems. *Trans. ASME-J. Basic Eng.*, **82**: 35–45.

[52] Welch G. and Bishop G. (1995) An introduction to the Kalman filter. Technical Report TR95-041, Department of Computer Science, University of North Carolina, Chapel Hill.

[53] Grewal M.S. and Andrews A.P. (2001) Kalman Filtering: Theory and Practice Using MATLAB, 2nd edn. John Wiley & Sons, NY.

[54] Cui Q. et al. (2005) Characterizing the dynamic connectivity between genes by variable parameter regression and Kalman filtering based on temporal gene expression data. *Bioinformatics*. **21**: 1538–1541.

[55] Alter O. et al. (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl Acad. Sci. USA*: **97**: 10101–10106.

[56] Nagy Z.B. et al. (2005) Real-time polymerase chain reaction-based exponential sample amplification for microarray gene expression profiling. *Anal Biochem*. **337**(1): 76-83.

[57] Kelemen J.Z. et al. (2006) Kalman filtering for disease-state estimation from microarray data. *Bioinformatics*. **22**(24): 3047-53.

[58] Armstrong S.A. et al. (2002) MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nat. Genet.* **30**: 41–47.

[59] Gordon G.J. et al. (2002) Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer Res.* **62**: 4963–4967.

[60] van't Veer L.J. et al. (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature.* **415**: 530–536.

[61] Yeoh E.J. et al. (2002) Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell.* **1**: 133–143.

[62] Fukunaga K. (1990) Introduction to Statistical Pattern Recognition (2nd edn). Academic Press, San Diego.

[63] Whitfield M.L. et al. (2002) Identification of Genes Periodically Expressed in the Human Cell Cycle and Their Expression in Tumors. *Molecular Biology of the Cell.* **13**: 1977–2000

[64] Troyanskaya O. et al. (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics.* **17**: 520–525.

[65] Joachims,T. (1999) Making large-scale SVM learning practical. In Schoelkopf,B., Burges,C. and Smola,A. (eds), Advances in Kernel Methods—Support Vector Learning. MIT Press, Boston, MA.

[66] Witten,I.H. and Frank,E. (1999) Data Mining: practical machine learning tools and techniques with java implementations. Morgan Kaufman, San Francisco, CA.

[67] Weston,J. et al. (2006) The Spider 1.71. http://www.kyb.tuebingen.mpg.de/bs/people/spider/main.html

LIST OF PUBLICATIONS

This dissertation is based on the following peer-reviewed publications:

[II]. **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.

[III]. 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.

[IV]. 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.

[V]. 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:

[VI]. 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.

ABSTRACT

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 *ntr*R 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 filtering 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.

ÖSSZEFOGLALÁS

Bevezetés

Az utóbbi években egyre gyarapodó biológiai adatbázisok – mint például a DNS szekvencia, génexpressziós mintázat, fehérje-fehérje kölcsönhatás adattárak – a rendszer biológia dinamikus fejlődését eredményezték. A ma hozzáférhető magas adatátvitelű gén-expressziós technológiák hasonlóan hozzájárultak a rendszer biológia tudományterület fejlődéséhez. A rendszer biológia lehetővé teszi az élettani folyamatok jobb megértését és az orvosi biológia területén megbízhatóbb diagnosztikát és orvosi kezelést ígér. Ez azáltal válik elérhetővé, hogy törekszik matematikailag modellezni és szimulálni a sejtben zajló komplex biológiai folyamatokat. Ismeretes, hogy a rákos megbetegedések altípusai eltérően válaszolhatnak az eltérő kezelésekre. Ezért is indokolt a kezelést megelőző pontos diagnózis. A gén-expressziós mintázata alapján, a rákos sejt egy olyan több-állapotos rendszerként fogható fel, ahol az egyes állapotok a rák altípusainak feleltethetők meg. Ez az elképzelés vezetett el a rákos megbetegedés gén-expresszión alapuló molekuláris klasszifikációjához – ami nem más, mint matematikai módszereken alapuló diagnózis. Az utóbbi években kitűntetett tudományos érdeklődésnek tartanak számot az elsősorban DNS microarray alapú ide sorolható módszerek. Nagyszámú mesterséges intelligencián alapuló algoritmusok, mint amilyenek a support vector machine, mesterséges neuron hálók, nearest neighbor osztályozó, vagy a random forests, azzal a céllal kerültek alkalmazásra, hogy pontosabb és megbízhatóbb diagnosztikát tegyenek lehetővé. Sajnos a meglévő gén-expressziós adatokon (QRT-PCR, DNS microarray), a kísérleti körülményekből adódó hiba (zaj) és a biológiai eredetű variancia együttesen megfigyelhető. Ezért indokolt egy több lépésből álló adat elő-feldolgozás és további módszertani fejlesztések is.

Célkitűzés

Célkitűzéseink a gén-expressziós adatfeldolgozással és módszertani fejlesztéssel kapcsolatosak. Nevezetesen:

- A standard gén-expressziós adatfeldolgozási módszerek alkalmazása QRT-PCR és microarray adatokon.

- Statisztikailag megvizsgálni, hogy a laboratóriumban használt protokollok alapján hogyan befolyásolhatók az egyes gén-expressziós változások.

- Klaszterezés és marker gének azonosítása szkizofréniás betegek gén-expressziós mintázatában.

- Új normalizációs és zajszűrési (Kálmán Szűrő) módszerek fejlesztése és alkalmazása a molekuláris szintű rák diagnosztikában.

A gén-expressziós kovariancia, mely a gének közti funkcionális kapcsolatot is mutatja, fontos szereppel bír a betegségek molekuláris osztályozásában. A Kálmán Szűrő figyelembe veszi a gén-expressziós kovarianciát. Célunk, hogy a Kálmán Szűrő segítségével kiszűrjük a kísérleti zajt és megbecsüljük a minták biológiai állapotát. Nem utolsó sorban szándékunkban állt megvizsgálni a Kálmán Szűrővel kezelt adatok osztályozhatóságát, osztályozó algoritmusok segítségével.

Eredmények

A disszertációban közölt eredmények bioinformatikai módszerek alkalmazását mutatják be. A kötelező gén-expressziós adat elő-feldolgozási lépések, nevezetesen a minőség ellenőrzés és a LOWESS normalizáció illetve a *t*-próba, mely a gén-expressziós eltéréseket tárja föl, a társszerzős publikációk eredményeiben kerültek bemutatásra. A fenti módszerek alkalmazásának részletes

leírása került bemutatásra, abban a publikációban, mely az ntrR által szabályozott géneket azonosítja Sinorhizobium meliloti modell organizmusban. Egy DNS microarray kísérletben az S. meliloti egy ntrR funkcióvesztéses mutánsát hasonlítottuk össze a vad típussal aerob és mikroaerob körülmények között.

Egerek makrofág sejtjein végzett lipopoliszacharidos kezelés egy olyan kísérletnek szolgált alapul, melyben a DNS amplifikációnak a gén-expressziós változásra mért hatását vizsgáltuk. A cDNS amplifikációt két protokoll - exponenciális fázisban megállított amplifikáció illetve szaturációs amplifikáció - alapján végeztük el és az eredményezett gén-expressziós változást mutató adatokon $\chi 2$ próbát hajtotunk végre. A kísérlet kontrolljaként egy non-amplifikációs protokoll szolgált. A statisztikai eredmények azt igazolták, hogy az exponenciális fázisban megállított amplifikáció megbízhatóbb, szemben a szaturációs amplifikációval, a microarray kísérletek reprodukálhatósága szempontjából.

Továbbá, egy szkizofréniás betegekből álló populációt használtunk fel arra, hogy megbízható marker géneket keressünk a kór molekuláris diagnosztizálásához. A DRD2 és a Kir2.3 bizonyultak marker génnek. Annak ellenőrzésére, hogy a fenti gének esetében valóban a betegség marker génjeivel állunk szemben, hierarchikus klaszterezést hajtottunk végre, beteg és egészséges személyektől származó adatokon. A klaszterező eljárás látványosan kimutatta, hogy a szkizofrén minták elkülönültek a normál mintáktól a fenti gének tekintetében.

A továbbiakban a gén-expressziós adatok klasszifikációja állt érdeklődésünk középpontjában. A klasszifikáció hatékonyságának javítása érdekében a Kálmán Szűrőt vezettük be. Munkánk szempontjából a legfontosabb tulajdonsága ennek a matematikai módszernek, hogy elkülöníti a biológiailag értelmezhető varianciát a mérési zajtól. A microarray kísérletben biológiai állapotnak tekintjük a gének valós expressziós szintjét. Az osztályozási felállásban ez az állapot az egyes alosztályoknak megfelelően változik. Ezt az esetet stochasztikusan modelleztük. A

mérési zaj szintén stochaszikusan volt megjeleníthető. A Kálmán Szűrő a stochaszikus modellek mellett fölhasznál még egy a microarray folyamatnak megfelelő determinisztikus modellt. Ezek segítségével vált felbecsülhetővé a gének valós expressziós szintje azaz a biológiai állapot. A fenti módszert 7 különböző publikus, tumoros eredetű adatsoron alkalmaztuk. A leghasználatosabb klasszifikációs módszereket szűrt és nem szűrt adatokon egyaránt teszteltük. Statisztikailag igazoltuk, hogy a Kálmán Szűrő szignifikánsan javítja az osztályozhatóságot. Három különböző grafikai ábrázolást alkalmaztunk, annak demonstrálására, hogy az egyes osztályok szemmel láthatóan elkülönülnek egymástól. Új markerek azonosítását is tárgyaljuk, annak bizonyítására, hogy a szűrt expressziós adatok, már kis számú gén esetében is predikciós portenciállal bírnak az osztályozásra nézve.

Notes sourced by www.wikipedia.org:

[i] **Alfred Bernhard Nobel** (1833-1896) was a Swedish chemist, engineer, innovator, armaments manufacturer and the inventor of dynamite. He owned Bofors, a major armaments manufacturer, which he had redirected from its previous role as an iron and steel mill. In his last will, he used his enormous fortune to institute the Nobel Prizes. There is no Nobel Prize for mathematics (the Fields Medal is often considered to be the equivalent in terms of prestige).

[ii] **Magnus Gustaf (Gösta) Mittag-Leffler** (1846-1927) was a Swedish mathematician. He was a member of the Royal Swedish Academy of Sciences (1883), the Finnish Society of Sciences and Letters (1878, later honorary member), the Royal Swedish Society of Sciences in Uppsala, the Royal Physiographic Society in Lund (1906) and about 30 foreign learned societies, including the Royal Society of London (1896) and Académie des sciences in Paris. He held honorary doctorates from the University of Oxford and several other universities.

[iii] **John Forbes Nash, Jr.** (1928-) is an American mathematician who works in game theory and differential geometry. He shared the 1994 Bank of Sweden Prize in Economic Sciences (also called the Nobel Prize in Economics)

[iv] **Daniel Kahneman** (1934-) is an American psychologist, notable for his pioneering work on behavioral finance and hedonic psychology.

[v] **Vernon Lomax Smith** (1927-) is professor of economics at George Mason University, a research scholar at George Mason's Interdisciplinary Center for Economic Science, and a Fellow of the Mercatus Center, all in Arlington, Virginia.

[vi] **Anatole France** (1844-1924) was the pen name of French author Jacques Anatole François Thibault. He was born in Paris, France, and died in Tours, Indre-et-Loire, France.

[vii] **Karl Ludwig von Bertalanffy** (1901-1972) was an Austrian-born biologist known as one of the founders of general systems theory.

[viii] **Anatol Rapoport** (1911-) is a Russian-born American Jewish mathematical psychologist. He is one of the founders of the general systems theory. He also contributed to mathematical biology and to the mathematical modeling of social interaction and stochastic models of contagion. He combined his mathematical expertise with psychological insights into the study of game theory and semantics. Rapoport extended these understandings into studies of psychological conflict, dealing with nuclear disarmament and international politics.

[ix] **Sir Arthur Stanley Eddington** (1882-1944) was an astrophysicist of the early 20th century. He is famous for his work regarding the Theory of Relativity. Eddington wrote an article in 1919, Report on the relativity theory of gravitation, which announced Einstein's theory of general relativity to the English-speaking world. Because of World War I, new developments in German science were not well known in England.

[x] **Francis Harry Compton Crick** (1916-2004) was an English molecular biologist, physicist, and neuroscientist, who is most noted for being one of the co-discoverers of the structure of the DNA molecule in 1953. He, James D. Watson, and Maurice Wilkins were jointly awarded the 1962 Nobel Prize for Physiology or Medicine "for their discoveries concerning the molecular structure of nucleic acids and its significance for information transfer in living material".

[xi] **Kary Banks Mullis** (1944-) is an American biochemist who developed the polymerase chain reaction (PCR), a central technique in biochemistry and molecular biology which allows the amplification of specified DNA sequences, for which he was awarded the Nobel Prize in Chemistry and the Japan Prize in 1993.

[xii] The *t*-statistic was introduced by **William Sealy Gosset** for cheaply monitoring the quality of beer brews. "Student" was his pen name. Gosset was a statistician for the Guinness brewery in Dublin, Ireland, and was hired due to Claude Guinness's innovative policy of recruiting the best graduates from Oxford and Cambridge to apply biochemistry and statistics to Guinness' industrial processes. Gosset published the t test in Biometrika in 1908, but was forced to use a pen name by his employer who regarded the fact that they were using statistics as a trade secret. In fact, Gosset's identity was unknown not only to fellow statisticians but to his employer—the company insisted on the pseudonym so that it could turn a blind eye to the breach of its rules.