Petri Törönen

**Analysis of gene expression data using clustering and functional classifications**

Doctoral dissertation

To be presented by permission of the faculty of Natural and Environmental Sciences of the University of Kuopio for public examination in Auditorium, MicroTeknia, University of Kuopio, on Saturday 6th November, at 12 noon
Abstract

Recently developed methods in biosciences enable simultaneous measurements of activity for thousands of genes. The bottleneck in analysis is typically the complexity of the resulting data. Moreover, the reliability of the obtained measures is often far from optimal. The current work presents new use of data clustering and visualization methods for analysis of gene expression data. Clustering provides a divide-and-conquer strategy to analysis whereas visualization aims at presentation of the data distribution that is useful for further understanding. We have mainly used Self-Organizing Maps as a starting point for clustering and visualization of the data.

After clustering and visualization methods we have analyzed which of the monitored functional gene classes were enriched in the obtained clusters. Enrichment means here that one observes more genes from a functional gene class than what would be expected by random in a cluster. Strong enrichment of the functional class increases the reliability of the association of the observed gene cluster with the cell function in question. This correlation also enables the sorting of the enriched classes to the estimated order of the importance for the data analysis and therefore simplifies the resulting analysis of the cluster area.

One goal was to automate the process of finding clusters with strong correlation with functional classes. This was realized by using hierarchical clustering as a starting point and selecting the clusters with the strongest correlation. The novel approaches developed in this thesis turned out to be useful enabling the planning of further experiments and creating new hypotheses. The presented methods could be used with other similar large data sets, as they could simplify the resulting analysis and uncover new details.

National Library of Medicine Classification: QU 26.5, QZ 50, QZ 52, WA 950
Medical Subject Headings: computational biology; gene expression profiling; oligonucleotide array sequence analysis; neural networks (computer); biometry; cluster analysis; classification; Saccharomyces cerevisiae; antipsychotic agents; brain; rats
Acknowledgements

This work was carried out in the A.I. Virtanen Institute for Molecular Sciences, University of Kuopio, during years 1999 – 2004.

I am grateful to my supervisors, Professor Eero Castren M.D., Ph.D. and Docent Garry Wong Ph.D. for the opportunity to work with bioinformatics and guidance. Garry has also contributed by reviewing my clumsy language.

Special thanks goes to the co-authors. These include Janne Nikkilä M.Sc., Markus Storvik Ph.D., Merja Oja M.Sc., Mikko Kolehmainen Ph.D., Juha Knuuttila M.Sc., Marketa Marvanová Ph.D., Sari Lähteinen Ph.D., Outi Kontkanen Ph.D.

Also, the official reviewers Docent Petri Auvinen Ph.D. and Professor Jaakko Hollmen Ph.D. improved significantly this work.

I wish to also thank professor Sami Kaski Ph.D. (Helsinki University) and his group members, Janne Nikkilä M.Sc., Janne Sinkkonen Ph.D. and Merja Oja M.Sc., for collaboration. They have also explained to me several issues from the field of data mining.

Many people have tolerated my endless questions in mathematics, statistics, programming and data mining. These include Lecturer Kari Mauranen M.Sc., Pirjo Halonen M.Sc., Professor Marko Vauhkonen Ph.D. and Olavi Manninen among others. A special thanks goes also to Comsol Finland for matlab language support and to Jarmo Pirhonen for visualization tips.

Thanks goes to my colleagues in Eero Castren’s group for bearing my existence. These include in addition to some of the previously mentioned also Mikko Sairanen M.Sc., Simon Beggs Ph.D., Eija Koponen M.Sc., Anna-Kaisa Haapasalo Ph.D., Tommi Saarelainen Ph.D., Topi Tervonen M.Sc., Panu Hendolin Ph.D.

Thanks to Niina for still living with me. Also thanks to Lola for keeping my reflexes in shape.

The Sigrid Juselius Foundation, Academy of Finland, Finnish Cultural Foundation and Kuopio University Foundation have financially supported this work.

Special thanks to all the rest that I have forgotten!
Abbreviations

A  adenine
BMU  best matching unit
C  cytosine
cDNA  complementary DNA
DNA  deoxyribonucleic acid
DTT  dithiothreitol
G  guanine
GO  gene ontology
ICA  independent component analysis
LLE  local linear embedding
MDS  multi-dimensional scaling
MIPS  Munich center for information on protein sequences
MK-801  dizocilpine
MM  mismatch
NAN  not a number
ORF  open reading frame
PCA  principal component analysis
PM  perfect match
RNA  ribonucleic acid
RoV  ratio of variance
SGD  saccharomyces genome database
SOM  Self Organizing Map
T  thymine
TS-SOM  tree-structured SOM
List of original publications

This thesis is based on the following publications referred to by their corresponding Roman numerals (I – V).


In addition some unpublished data is presented. Names of the authors appear with different letters in some of the articles (for example Törönen or Toronen) due the differences between Finnish and English alphabets.
# Table of Contents

1. Introduction .......................................................................................................................... 3  

2. Review of the literature ......................................................................................................... 3  
   2.1. Methods for measuring gene expression data ................................................................. 4  
      2.1.1. cDNA glass chip .................................................................................................. 5  
      2.1.2. Affymetrix oligo chip ....................................................................................... 5  
      2.1.3. Differences of Affymetrix and cDNA chip method ......................................... 5  
      2.1.4. cDNA macroarrays ......................................................................................... 6  
   2.2. Obtaining and processing gene expression data .............................................................. 6  
      2.2.1. Image Analysis .................................................................................................. 6  
      2.2.2. Quality measures for image analysis .................................................................... 7  
      2.2.3. Data preprocessing ............................................................................................ 8  
         2.2.3.1. Normalization of measurement groups ........................................................ 8  
         2.2.3.2. Measure of differential expression .......................................................... 11  
         2.2.3.3. Measures for replicated data ........................................................................ 13  
         2.2.3.4. Similarity measures (distance metrics) ........................................................ 14  
   2.3. Analysis of the data ......................................................................................................... 16  
      2.3.1. Identification of regulated genes ..................................................................... 16  
      2.3.2. Dimension reduction methods ............................................................................ 17  
      2.3.3. Clustering of the data ...................................................................................... 18  
         2.3.3.1. Self-Organizing Maps (SOM) ...................................................................... 20  
         2.3.3.2. Tree-Structure SOM ..................................................................................... 21  
         2.3.3.3. Methods similar to SOM .............................................................................. 22  
         2.3.3.4. Visualization of the SOM result .................................................................. 22  
      2.3.4. Evaluation of the clustering results ....................................................................... 22  
      2.3.5. External information for the evaluation of the clusters .................................. 23  

3. Aims of the study ................................................................................................................ 24  

4. Materials and methods ........................................................................................................ 24  
   4.1. Data sets ..................................................................................................................... 24  
      4.1.1. Publicly available data sets ............................................................................. 24  
      4.1.2. In house produced data sets ............................................................................ 26  
   4.2. Preprocessing ............................................................................................................. 27  
      4.2.1. Preprocessing of public data sets .................................................................... 27  
      4.2.1.1. Novel method for filtering low variance genes ............................................ 28  
      4.2.2. Preprocessing of in house data sets ................................................................. 30  
   4.3. Algorithms .................................................................................................................. 30  
      4.3.1. Clustering algorithms ...................................................................................... 31  
      4.3.1.1. Self-Organizing Maps .................................................................................. 31  
      4.3.1.2. Visualization and clustering of SOM results ............................................... 31  
      4.3.1.3. Hierarchical Clustering ................................................................................ 31  
   4.4. Gene annotations and classifications ............................................................................ 32  
      4.4.1. Correlation measure ........................................................................................ 32  
      4.4.2. Applications of the correlation measure .......................................................... 36  
      4.4.3. Searching method for clusters enriching functional classes for cluster trees . 37  

5. Results .................................................................................................................................. 39
1. Introduction

Recently developed methods in the biosciences enable simultaneous measurements of activity for thousands of genes. These methods have allowed a new level of analysis of biological and pathological phenomena where, instead of the activity of a few genes, the emphasis is on analyzing the genes important for the phenomena. This helps to create an almost complete picture of the cellular activities in the analyzed biological samples. The applications range for example between analysis of pathological samples in cancer (Alizadeh, et al., 2000, Alon, et al., 1999), analysis of single cell organisms in stress treatments (DeRisi, et al., 1997), to comparison of mutation strains of the organism to normal strains (Hughes, et al., 2000). These methods have currently become a standard in biosciences.

The drawback is the size and the complexity of the obtained data, making any manual analysis often impossible. Also, the reliability of the obtained measures is often far from optimal, which combined with the large number of measurements increases the probability that some of the observed gene activities are not reliable. If these drawbacks are not considered in the data analysis the results might lead to erroneous conclusions.

In order to solve some of these associated problems, the current work presents use of data clustering and visualization methods for analysis of gene expression data. Clustering creates subgroups from the data that can be analyzed separately therefore simplifying the massive analysis task. Visualization aims at simplified presentation of the data that preserves the most information from the original distribution. We have mainly used Self-Organizing Maps as a starting point for clustering and visualization of the data.

Despite the clustering step, the analysis of the obtained clusters often requires vast information on the analyzed genes. Therefore we have utilized a statistical method, Fisher exact test, with available categorization of genes to different functional groups after clustering. Fisher exact test looks for the analyzed functional gene classes showing significant correlation with the obtained cluster areas. This correlation is based on observing more genes in a cluster from functional gene class than what would be expected by random. Enrichment of the functional class increases the reliability of the association of the gene cluster with the cell function in question. The results from Fisher exact test can be also used to sort the correlating classes to the estimated order of the importance for the data analysis. This is shown to simplify the resulting data analysis.

The information presented by the clusters showing a strong correlation with functional classes turned out to be useful. This created an interest to directly look for the clusters with strong correlation to functional classes. Two methods are being presented for this purpose.

2. Review of the literature

One of the important advances in the biosciences during the last ten years has been the development of methods for the analysis of gene expression and especially the
development of the DNA microarray technologies. Miniaturization of the analysis platform enables the monitoring of expression level of thousands or tens of thousands of genes at the same time (DeRisi, et al., 1997, Schena, 1996). These observed gene levels could be used to create a ‘finger print’ for the state at which the cells at the monitored sample reside. One could, for example, notice that the genes associated with an inflammatory response are active at the area of infection or that different pathways are activated in microbial organisms when different carbon sources are available. The obtained gene expression fingerprint could also be used to associate active unknown genes to the analyzed cellular state. Similarly, the fingerprint could be used to recognize the cellular state when the information about the cellular state is missing. This can be done using known active genes in the sample and/or comparing the obtained gene expression fingerprint of an unknown sample to the known samples. A similar fingerprint can be created to each gene with the profile that shows gene activity in different samples. These gene specific gene expression profiles could be used likewise to identify the function of unknown genes by looking at similarities in expression with genes having a known function.

The measurement of gene expression is currently done mainly with chip technologies. These chips include areas that each bind selectively expressed RNA of one gene. The bound RNA is labeled with fluorescence or radioactive labels. The signal from these labels can be quantitated using image analysis methods. The obtained measure and its estimated error can be used for the analysis of the genes presented on the chip. The following chapters explain these steps in more detail.

2.1. Methods for measuring gene expression data

The basis of gene expression measurements is the use of probe DNA sequences that bind specifically to RNA or DNA sequences of interest. The probe DNA sequences are usually attached to a surface such as nylon membrane or glass. DNA sequences are constructed with four types of nucleotides: adenine (A), guanine (G), cytosine (C), and thymine (T). Each of these nucleotide types binds strongest with the corresponding nucleotide type (G with C, A with T and vice versa) and a sequence of nucleotides equally binds strongest the corresponding sequence. Such corresponding nucleotide sequence can be used to bind messenger RNA of interest from sample. After binding the sequence of interest the other unbound nucleotide sequences can be washed from the surface and the amount of bound nucleotide can be measured.

Probe sequences can be complementary DNA (cDNA). These are equally long as the analyzed sequence matching it over the whole sequence length. Probe sequences can also be shorter oligonucleotide sequences that match only to a selected region of the analyzed sequence. Usually the selected region is chosen so that it would bind optimally the analyzed messenger RNA with a minimum amount of binding to other sequences in the pool of all possible RNAs in the cell. It is customary to use several different oligonucleotides for analysis of one gene to improve the result. Measuring the amount of RNA has been a standard technique in biochemistry but the major step in the recent development has been the strong parallelization that enables measuring thousands of genes from the same sample. Only a short description of these methods that have become
standard in biosciences is given here. Readers not familiar or highly interested are recommended to see vast the literature on the area (Duyk, 2002, Johnston, 1998, Marvanova, 2003, Schena, 1996).

2.1.1. cDNA glass chip

One version of these chips is a chemically modified glass slide to which each of the several thousands of probes has been attached in an array format. Probe sequences are transferred in a parallel fashion using a robotic arm to glass slide chemically treated to bind nucleotides (Shalon, et al., 1996). The probe is usually cDNA but also synthesized oligosequences of different lengths have been used (Hughes, et al., 2001). After the probes are attached the glass slide has areas including different probe sequences (so called spot areas) in an array format. Each probe gives a measurement only for a single RNA but the array construction enables the parallel analysis of thousands of genes from the same sample. Analysis is usually done using two samples at the same time. These samples are labeled with different fluorescence dyes and usually one of these samples is a control sample that works as a reference and the actual test sample measurement is compared against this. There have been proposals of obtaining more than two measurements from one spot so that, for example, three samples would be compared or that one measurement tests the amount of probe DNA in the spot area (Hessner, et al., 2003).

2.1.2. Affymetrix oligo chip

Another popular version of the chip designs is the one of Affymetrix where lithographic technologies familiar from computer silicon chip manufacturing are used to synthesize oligonucleotide probes on a silicon chip (Lockhart, et al., 1996). Oligonucleotides in this design are 25 nucleobases long. With the Affymetrix method, each analyzed gene is monitored usually with several (~11 - 20) different oligo probes each having a perfect match (PM) to the sequence of interest at different selected areas of sequence. At the same time the background for unspecific binding to sequences and background signal for image analysis is estimated with similar oligo probes having one wrong nucleotide at the middle of the sequence (mismatch, MM). With Affymetrix chips each sample is measured separately and no reference sample is included. There has been criticism about the use of MM in the data analysis (Irizarry, et al., 2003). This probably results from too small number of mismatch nucleotides in the MM probe which causes it to measure too much the actual correct sequence (Hughes, et al., 2001).

2.1.3. Differences of Affymetrix and cDNA chip method

Differences between the two methods are that in cDNA chips it is possible to compare two samples against each other. This should cancel some of the error sources from the measurements like weaker signal from one spot than from some other spot. On the other hand, the Affymetrix chips have parallel measurements of the amount of the analyzed genes. Parallel measurements have been used in advanced data analysis methods to
discard unreliable measurements and to increase the reliability of the observed signal (Comander, et al., 2004, Irizarry, et al., 2003, Lemon, et al., 2003, Li and Wong, 2001). Also another benefit in the Affymetrix chip design is the control for the unspecific binding with MM oligo probes although this has been criticized as was pointed out previously. These can be used to control the error caused by binding of similar sequences to the probes. Also an important difference between methods is that the manufacturing of custom glass arrays is usually possible for large research institutions and universities whereas the manufacturing of the Affymetrix type chips would require advanced technology not feasible for research units.

2.1.4. cDNA macroarrays

DNA filters (see (Southern, et al., 1999) for tutorial) form the third method for producing gene expression data and it is older than two other methods. cDNA macro arrays have some common features between both of the previous methods. These usually include only one probe for each analyzed sequence and usually this probe is cDNA making them similar to cDNA chips. The actual sample is placed over the whole area of the filter and only one sample is analyzed with each filter making these filters similar to Affymetrix chips. Radioactive labels are used instead of fluorescence labels utilized in previous methods.

2.2. Obtaining and processing gene expression data

2.2.1. Image Analysis

The first step in the production and the analysis of the gene expression data is the image analysis. Image analysis aims at quantitating the bound cDNA from the area of each probe sequence. The quantitation step is important for the data analysis as the methods that are robust to error signals in the obtained image enable more reliable analysis. Also the errors in this quantitation are hard to correct later in the analysis. Due to the importance of image analysis to the later steps in the preprocessing, it has been included here as a part of the data processing steps.

The bound cDNA is observed using fluorescence labels attached to the cDNA in the analyzed sample. Radioactive labels have been also used when analyzing cDNA macro arrays. These are in turn quantitated for each pixel area. Actual image analysis is usually based on the obtained mean or median pixel signal for the obtained probe area (often referred as a spot) and the same value for the possible background area that is used to estimate the background signal. The background area signal is normally extracted from the actual signal (Eisen, 1999). The obtained results are usually compared to a reference sample by taking the ratio between sample result and reference result (Chen, et al., 1997, Shalon, et al., 1996). A similar step can be done in larger analysis sets with Affymetrix chips by comparing the obtained result of each gene with the average of all the measurements for the gene. All data sets analyzed here are based on this or a similar image processing step.
As an alternative for analyzing the median or mean signal from the spot area some authors have proposed the pixel level ratio analysis where separate pixels are compared between sample and reference and the median or average of these values is used (Bozinov and Rahnenfuhrer, 2002, Brody, et al., 2002, Eisen, 1999, Rahnenfuhrer and Bozinov, 2004). Analysis of pixel level ratio information aims at lessening the effect of pixels with outlier results to the obtained result. It has been shown for example that the distribution of results obtained using median of the pixel level log of ratios is closer to normal distribution than the log ratio of medians (Brody, et al., 2002). Log ratio of medians was actually shown to form a Cauchy distribution that is known for not having a defined expectation value. It should be also noted that the pixel level information can be used to correct some errors in the data like the saturation of the pixel measurements (Dodd, et al., 2004).

### 2.2.2. Quality measures for image analysis

Reliability of the obtained results is always a key question for any data analysis. Already the first program specially created for cDNA chip image analysis, Scanalyze (Eisen, 1999), included evaluation of correlation of the pixel level information between the sample and the reference channel. The importance of the error measures can be thought with simple example. Data can include two series of measurements where one of them shows ten times bigger variance. If no error measure were available it would be hard to tell whether the bigger variance is caused by stronger signal in one group or is it just a result of the bigger amount of error in the signal.

Data analysis methods aim at filtering out two different types of error from the obtained measurements: one is additive error like background or zero level variation obtained at different areas of the chip or filter; the other is multiplicative error, like the affinity of the specific probe sequence to target sequence or the efficiency of the used fluorescence dye. The difference between additive and multiplicative error is that the additive is independent from the signal level while the latter depends on the magnitude of the signal. Errors in the measurements play an important role throughout the analysis of the data. More detailed analysis of the possible sources of errors have been described (Dror, et al., 2003, Schuchhardt, et al., 2000).

One of the evaluation methods is the variance of pixel level information for the quantitated spot (Brown, et al., 2001). This is a natural selection as it quantitates how similar the result stays over the area of the spot and it is also associated with the use of pixel level log-ratio measurements for the quantitation (see previous chapter). Also the uncertainty of the spot placement has been used in the evaluation by testing how the log of the ratio of the mean values varies when the uncertainty about the spot placement is taken into consideration (Lawrence, et al., 2004). A combination of several error quantity measures has been also used (Wang, et al., 2001). It should be noted that most of these image analysis level error measures seem to be able to monitor both the additive and multiplicative error at the same time by showing bigger errors for the obtained ratio values among the low intensity measurements (see for example (Brown, et al., 2001,
Lawrence, et al., 2004, Wang, et al., 2001)). One data set used in this work (Hughes, et al., 2000) included an error model that was based on the estimates of additional and multiplicative error in the obtained measures. These errors were estimated using data from repetitive analysis of control samples. Also other similar expression measure error models have been proposed (Dror, et al., 2003, Rocke and Durbin, 2001).

Error measures can be used throughout the later analysis, for example, to scale the importance of different observations in the distance measure (Hughes, et al., 2000). The obtained error can be also used to weight the replicated measurements with the inverse of the error when combining the replicates (Brown, et al., 2000, Hughes, et al., 2000). The use of error ensures that the bad quality replicates do not effect the obtained result as much as good quality replicates. One other possibility is to confirm the results in the later analysis by doing repetitively the analysis steps with measures that are sampled using the error distribution (Li and Hung Wong, 2001). Estimated quality of the spot area can be also used to discard less reliable measurements (Hautaniemi, et al., 2003). Repeated analysis can be used to test how robust the obtained results are against the error in the data set.

2.2.3. Data preprocessing

In the preprocessing steps the aim is to increase the reliability of the observed measurements and to cancel the effects of the previous error signals. The chapter concentrates mainly on the analysis of data produced with cDNA microarrays. Preprocessing steps and their effects have been described in detail (Cui, et al., 2003). Preprocessing steps have been evaluated also using a control data produced with known amounts of RNAs (He, et al., 2003).

2.2.3.1. Normalization of measurement groups

Additive and multiplicative errors usually have components that can be estimated and components that cannot be known. Components that can be estimated are often on the same level, especially for the genes coming from the same set of measurements. Systematical additive error that can be estimated will be referred to as a bias and systematical multiplicative error that can be estimated will be called multiplicative bias. The extraction of these biases from the data groups is called normalization. “Group” here can refer to genes coming from the same chip or it can also be a part of the measurements from the chip like the data coming from one sub-grid of the cDNA chip (Dudoit, et al., 2000). These groups of data should be selected so that they can be expected to have similarities in additive and/or in multiplicative bias. Although biases can be taken care of it should be noted that the data still certainly includes error components that cannot be estimated.

The main assumption in most normalization procedures is that most of the genes stay approximately at a constant level. Although this assumption normally holds, there are test situations that can violate this assumption. In those cases one should use housekeeping
genes or control RNA sequences that are added to the test material. Housekeeping genes refer to genes that are believed to stay at a constant level despite the cellular changes occurring through the test situation. Control RNAs on the other hand are RNA sequences not found in the sample. A constant amount of these is added to each analysis sample. Both groups work as a reference point for the rest of the measurements.

A standard procedure for correcting the bias in the measurements is the extraction of the average or median signal level from the measurements. The median value is usually considered a better choice (He, et al., 2003) as it is less sensitive to outliers. Such extraction would take care of additive bias. Multiplicative bias can be taken care of similarly by looking at the average or median of the ratio between measurements in separate measurement groups and correcting it similarly to one. These steps can be also combined using for example least squares regression that combines these two steps (see examples in (Cui, et al., 2003)).

The gene expression data has been observed to contain many other analysis disturbing artifacts. It was observed early in Dudoit et al (Dudoit, et al., 2000) that the average of the intensities from control and sample affect the average ratio. Figure 1 shows the first time point from DeRisi et al. (DeRisi, et al., 1997) demonstrating this artifact. There are many proposed solutions also to this artifact that are based either on the extraction of the local bias from the signal (locally weighted least squares regression, lowess method) (Dudoit, et al., 2000) or optimization of the logarithm function where bias is corrected by including a correcting constant to the ratio equation (Sapir and Churchill, 2000). Lowess method has been also used in our work with in house data (Knuuttila, et al., 2004)
Figure 1. Visualization of biases in the expression data with no biological signal.

The figure shows an example of the resulting signal distribution when the same sample is compared against itself (first time point from DeRisi data (DeRisi, et al., 1997) explained in Methods). Note that the ‘correct’ distribution would be that all the data points would lie on the $y = 0$ line, or close to it, which would signal that the ratio stays constant. Notice also the drift upwards in the signal among the weaker signal levels. Drift shows that the average signal contains an intensity dependent bias. A similar presentation is shown in the fig. 2 b where the intensity dependent differences in variance are shown.

As the normalization procedures are based on the analysis of genes with approximately constant expression in the expression data, it is reasonable to discard genes with strong change in the expression from the normalization procedure (outlier genes). Those genes that stay approximately at the same level in the normalized data will have little difference in their rank values in two data sets. Based on this, the selection of the suitable genes for the normalization procedures has been done using the ranks of genes from two normalized data sets so that only the genes with small change in their rank value were used for the normalization procedure (Tseng, et al., 2001).

The danger in the normalization procedures is that the data set is normalized too much causing the normalization steps to filter the actual signal from the data set. As an example, one could imagine a situation where most of the genes are down regulated. If the normalization of this data set were based on the average expression profile it would
make close to a constant expression profile for the down-regulated genes. Such situations require the earlier mentioned use of spiked in control genes or housekeeping genes. Before the actual analysis the proportion of the regulated genes should be estimated. For example the treatment of a cell line with close to lethal amount of chemical compound will surely cause drastic responses in gene expression whereas normal developmental changes in cells often regulate only a small proportion of genes. Note still that even when the standard normalization procedures can be used the spiked in controls and housekeeping genes can still be used to evaluate the obtained normalization result.

2.2.3.2. Measure of differential expression

It is customary to take a ratio between the sample of interest and the reference sample. This analysis will cancel out the multiplicative biases that are equal size in the sample of interest and in the reference. This is very natural especially when analyzing the data using the cDNA arrays, as the sample and the reference have been measured using the same spot area. So the multiplicative biases associated with the spot area should cancel out. The log-value of the obtained ratio value is currently the most popular measure for the analysis of up or down regulation. The drawback of this measure is that it does not take into consideration the additive error in the data. Additive error has been observed to cause increase of variance among the small intensity measurements (Durbin, et al., 2002, Huber, et al., 2002, Hughes, et al., 2000, Rocke and Durbin, 2001). We have also observed the same phenomena in our own data (see fig. 2).
Figure 2. Effect of the constant in the ratio equation to the obtained results and the visualization of large variance among the weak intensities.

Figure 2. A shows the scatter plot of the measurements (sample vs. control). Notice that when the normal ratio is calculated (2. B) the variance among the small signal intensities is extremely large. This artifact can be corrected by using the corrected ratio (Newton, et al., 2001) (Durbin and Rocke, 2004) as is shown in fig. 2 C. Notice that the wide spread of small measurement intensities has disappeared. X-axis in 2 B and 2 C are different from fig. 1 as here the x-axis shows the average of original intensities and in fig. 1 the x-
There are several corrections to this problem. One natural option would be to monitor both the difference and ratio between sample of interest and reference sample (Cole, et al., 2003). The problem with the analysis of both the difference and the ratio is the question on how to combine these two measures into one measure. Log-ratio measurements have been also scaled with estimated error as proposed in the image analysis part (Brown, et al., 2001, Hughes, et al., 2000, Lawrence, et al., 2004, Wang, et al., 2001). Variance stabilizing functions have been also proposed instead of logarithm (Durbin, et al., 2002, Durbin and Rocke, 2004, Huber, et al., 2002). Variance stabilizing functions are functions that diminish the variance among the low intensity measurements. As the additive error causes variance to increase among the measurements of small size one way would be to scale the obtained variance of log-ratio measurements with the local variance of data (Kepler, et al., 2002).

Data has also been fitted with statistical models like independent gamma distributions (Newton, et al., 2001) to raw measurements or strongly correlating normal distributions to log transformed values (Loguinov, et al., 2004). Also these methods enable the minimization of the effect of the additive error to the analysis of the data.

Later chapters (4.2) show the use of two methods to solve this problem. One of the public data sets included an error estimate for the log-ratio (Hughes, et al., 2000). This was used in the analysis to filter the effect of different errors by scaling the log-ratio with the estimated error. The reported error from the image analysis step could be used similarly here. With in house data sets, a simple method was used where a constant was added to the numerator and denominator in the ratio calculus (Durbin and Rocke, 2004). Use of such constant diminished the variance among the low-intensity measurements (compare figure 2 B and 2 C).

### 2.2.3.3. Measures for replicated data

One natural way to increase the reliability is to make repetitions for the measurements. Obtained results will give an estimate about the average value for log-ratio measurement and variance for the log-ratio measurement. These can be used with statistical tests to evaluate the significance of two groups when the variance between the repetitions is also taken into consideration. Analysis can be done using only the averages, but also a statistical test called t-test is available for this type of testing (Dudoit, et al., 2000). T-test can be used to answer the question: what is the probability that the obtained result or more extreme could be observed between the measurements coming from the same group.

The drawback is that due to the usual small number of repetitions normal t-test is not reliable. The measurements can show insignificant difference between two measurement groups but if the variance happens to be even smaller, the result will be considered significant. Another opposite example is the situation where there is an enormous amount
of variance in the measurements but the sample signal is all the time stronger than the control signal. This would be considered less significant.

The main problem with a small number of repetitions is that the obtained measure for variance is not reliable (Baldi and Long, 2001, Lonnstedt and Speed, 2002). Measurements can be created by a distribution of fixed variance but when the number of replicates is small the obtained estimate of variance can vary a lot from the correct value of variance. If the estimate is too small it can make the obtained insignificant result seem significant and vice versa with too big an estimate. The problem is increased by the big number of parallel analysis of many genes with little number of replicates as there is always certainly some genes in the gene pool that get a small variance just by chance.

One way to optimize the t-test for small repetitions is by including ‘a priori’ information about the variance of the genes into the t-test equation (Baldi and Long, 2001). A priori information is often used in bayesian data analysis, and here, a priori information works as a starting point assumption about the variance of the genes that can be perfected with the observed variance. The local average of variance for the genes with similar size of average of sample and control signal has been used for example as the initial estimate (Baldi and Long, 2001). The used estimate was able to take into account the bigger expected variance among the low intensity measurements. Also another similar method that applies bayesian statistical analysis has been presented (Lonnstedt and Speed). Both these resulting tests are roughly in between the testing only with average and testing with t-test as they lessen the effect of the observed variance.

2.2.3.4. Similarity measures (distance metrics)

Many of the later analysis steps for multi-dimensional data require the definition of the measure for the similarity of two data vectors (distance metric or similarity measure). Similarity measures define what is considered important when comparing multiple measurements for analyzed genes (gene data vectors). Similarity measure is needed to select the genes with the most similar measurements when predicting the function for gene with no known function. It is also needed when trying to group genes into groups of genes with similar expression in the experiments.

As an example, the same size of measurements could be considered important while comparing data vectors. Typically, for this kind of purpose, the Euclidean metric has been used (fig. 3 A). Optionally, a similar shape of the data vectors could be considered important even though the actual measurements would differ a lot (see fig. 3 B). In this case, the variance of the data vectors could be scaled to the same value before use of the Euclidean metric. After such scaling, the differences in the variances of the data vectors will not affect the analysis anymore. Optionally, the average of the data vector can be also scaled to the same value causing the differences in the average not to affect the analysis anymore (see fig. 3 C).
Figure 3. Behavior of different similarity measures

The figure shows how the preprocessing steps, like the normalization of variance of the expression profiles or normalization of average of the expression profiles, affect the similarities of the expression profiles. White and black symbols show the resulting groups after the preprocessing. Expression profiles and the resulting groups are different in each case to highlight the different features each resulting distance measure considers important. (A) Euclidean metric. Here the data is used for distance analysis with no preprocessing steps. Two clusters show the same level of expression but the resulting expression profile shapes are different. (B) Variance normalized metric. The variance of two expression profiles is scaled to 1. Results show two groups with strong similarity with the same shape but with a different average. (C) Average normalized metric. The average of the expression profiles is scaled to 0. Resulting groups appear after moving expression profiles to the same level. (D) Normalization of the average and variance. All the
It is normal to combine the two previous steps so that the average of the data vector is first scaled to null and the variance of the data vector is scaled to one after that (see the fig. 3 D). The resulting distance metric is similar to Pearson correlation, which is one of the most popular distance metrics. Pearson correlation measures the linear correlation between the measurements in two data vectors. Several other similarity measures have been used in the analysis of gene expression data, but usually most of them can be thought as variations from Euclidean metric or from Pearson correlation with the exceptions like mutual information of the data vectors (Steuer, et al., 2002) or pattern finding directed metric (Lepre, et al., 2004). More details about the various distance measures is presented in the literature (Toronen, et al., 2003).

It is important to notice that the scaling of the variance will multiply the variance of those data vectors that have insignificant variance. These data vectors usually have only signal that is caused by the error in the measurements and once the variance is scaled to one these cannot be told apart from the data vectors that really have a meaningful signal. This is why a filtering step where data vectors having too small variance are usually included in the analysis steps when Pearson correlation or scaling of the variance is used. A novel and simple method for this step is discussed later (see chapter 4.2.1.1).

As there are many similarity measures and other preprocessing steps to select from, the question arises: which preprocessing steps one should choose? Oja and coworkers presented one solution to this question by comparing how well the preprocessing steps grouped together genes with the same function (Oja, et al., 2002). A similar method is also presented elsewhere in the literature (Gibbons and Roth, 2002). A new method that can be applied to this evaluation is later presented (see chapter 4.4.3).

2.3. Analysis of the data

Two main approaches can be defined among different ways of analyzing the obtained data: hypothesis driven and exploratory. A hypothesis driven approach would start with a question like: are the ribosomal proteins associated with the cell cycle? Analysis methods would be then used to test whether the hypothesis is true or false. Exploratory analysis starts with no hypothesis. The main emphasis is in viewing data so that the analyst could obtain some preliminary hypotheses. The obtained hypotheses can be later confirmed with hypothesis driven data analysis. The emphasis in this work is on the exploratory analysis step that forms usually the first step in the analysis.

2.3.1. Identification of regulated genes

The simplest type of analysis for gene expression data is the selection of regulated genes for one treatment. Such work aims at separating the up-regulated, the down-regulated,
and the genes with no regulation at all from each other. Although the previous question of regulation is simple, problems arise while trying to create reliable number of replicates under the limiting factors like available amount of sample or the increasing costs for the research. All the previously described methods (see chapters 2.2.3.2. and 2.2.3.3.) that increase the reliability of the observed change in the expression have been used for this purpose. Besides the simple analysis of genes that were observed to be up or down-regulated also the groups of genes have been viewed similarly (Mirnics, et al., 2000) where different functional groups were monitored for the regulation. This method benefits from the monitoring of several genes at the same time causing the errors associated with the measurements of one gene to effect less on the obtained result.

2.3.2. Dimension reduction methods

Analysis of up or down-regulation using normal simple scatter plots can be adequate while analyzing one to three different treatments, but after more treatments, it starts to become increasingly harder to create an informative presentation of the data. Each split of data to up-regulated, down-regulated and unregulated genes also multiplies the number of resulting gene groups by three. One way to analyze a larger data set would be the visualization of the data using multiple scatter plots each showing one pair of treatments for genes. This can be useful especially with a different coloring for interesting groups of genes. Still the drawback is the large number of resulting visualizations.

Dimension reduction methods are usually a more useful way to visualize large-dimensional data. These methods aim at reducing dimensions so that the amount of the information lost in the transformation would be minimal. Usually dimension reduction is based on the similarity of the measurements (such as similar tissue samples or close time points in the time series), which can be presented by one variable using previously described distance measures.

There are two groups of methods for the dimension reduction, linear and non-linear methods. Linear methods, like Principal Component Analysis (PCA) and Independent Component Analysis (ICA), aim at creating a linear presentation of the original data. It can be considered as selecting a new set of axis for presentation of the data by drawing straight lines through the data space for new reduced dimensionality. Roughly, these lines are selected in linear methods so that the most of the information of the data points is conserved in the resulting space. This is done with PCA by looking at independent components that best preserve the variance in the original data. ICA functions similarly as PCA but it emphasizes the non-gaussian distributions along the extracted axes.

The other main group is non-linear dimension reduction methods. These mainly aim at preserving the same genes in the neighborhood of each gene from the original data in the resulting visualization. This can be imagined as a variation of the previous example where straight lines were drawn through data space. Non-linear methods (such as Multi-Dimensional Scaling (MDS), Isomap (Roweis and Saul, 2000), Local Linear Embedding (Tenenbaum, et al., 2000), Sammon’s mapping (Sammon, 1969), VxInsight (Davidson, et
allow these lines to curve and they can be stretched or shrinked in order to allow optimal separation of the data points. Preservation of the neighborhoods is expected to be more useful while using the resulting visualization for example to associate the function of neighboring known genes with genes of no known function. The drawback is that some non-linear methods usually require more user-defined parameters (Isomap, LLE) and/or need more optimization for the end result (Sammon’s mapping and MDS for example). Later work (chapter 4.3.1.2) demonstrates one way of simplifying the optimization with non-linear dimension reduction method. Many well known dimension reduction algorithms have been compared by viewing how well they conserve the neighborhood of each data point in the resulting visualization (Kaski, et al., 2003) which can be used to evaluate how useful the resulting visualization is for the analysis of data. Still even the best performing dimension reduction methods will distort the neighborhoods of data points. The amount of lost information usually increases as an increasing number of dimensions is being compressed.

### 2.3.3. Clustering of the data

Another solution for high-dimensional data is to cluster analyzed data points (genes or samples) using the expression profile. Clustering aims at aiding the analysis by applying the divide-and-conquer principle. Clustering allows the analysis of a smaller group of genes with similar expression profiles. At the same time the genes observed with a similar expression profile increase the support that the observed expression profile of the cluster was not just an artifact caused by error signals in the single gene expressions. Clustering is a vast part of data analysis and a more detailed analysis of clustering methods can be found from literature (Jain AK, 1988, Toronen, et al., 2003).

The first clustering methods applied to gene expression data were hierarchical clustering methods (Eisen, et al., 1998). Hierarchical clustering methods either join clusters or data points in a hierarchical fashion or split clusters hierarchically. The methods that join the clusters (agglomerative methods) usually start from single data points, join repetitively by selecting the most similar pair, and stop when all the data forms one single cluster. Exceptions from the previous procedure include hierarchical clustering of clusters created with preliminary clustering (Herrero and Dopazo, 2002). Methods that split data (dividing methods (Alon, et al., 1999, Herrero, et al., 2001)) usually start from the whole data as one cluster, split it into smaller clusters by selecting always the most dissimilar pair of resulting clusters and stop when each cluster includes only one data point. A suitable stopping criterion could be used to select a level after which the obtained clusters are not split any more (Herrero, et al., 2001). Hierarchical clustering methods can be used to create a tree for visualization of the data (Eisen, et al., 1998). This visualization could be cut at a selected level creating a grouping of the data points to resulting clusters.

Three hierarchical agglomerative clustering methods have been used regularly with gene expression data. These methods are: complete, average linkage, and Ward’s minimum variance method. The only difference between the methods is the way the distance of two joined clusters is defined. The complete method takes all the pair-wise distances from
members of one cluster to members of another and selects the biggest pair-wise distance as the distance of two clusters. Average linkage method takes the same group of pair-wise distances as the previous method but calculates the average linkage of all the pair-wise distances and takes that as the distance of two clusters. Ward’s minimum variance method is similar to average linkage method but instead of calculating the mean of the group of pair-wise distances the mean is calculated with squared pair-wise distances. This guides the clusters towards minimum inner-cluster variance. When these three methods are being compared, one notices that the summing of squared distances makes the resulting distance more sensitive to bigger pair-wise distances when comparing the clusters. So, Ward’s method’s results should be more similar with the results of the two other methods than what the results from average and complete linkage are (see chapter 5.2 for comparison). In addition, a fourth regularly used agglomerative method is the centroid method. This method defines the distance of clusters as a distance between the average center points of the clusters.

Another popular method for clustering gene expression data is k-means clustering (Jain AK, 1988, Tavazoie, et al., 1999). K-means starts with the number of clusters defined by user. Next the starting points for center points of these clusters are selected (often by random). Next the clusters are allowed to move in the data so that each data point is selected one at the time to pull the closest cluster center towards itself. This is repeated iteratively lessening the amount of pull in each repetition. After the algorithm has run selected number of rounds, the data is clustered to closest cluster center. The pull in the k-means should result in an optimal grouping of the data to clusters (clusters have data points inside the cluster as close to the center point as possible while at the same time having a maximal distance to members of other clusters). Another option to perform k-means is to select the cluster center starting points and place data points into the clusters similarly as before. After this the average position of the data points closest to each cluster center can be calculated. The obtained average position is then selected as the new position for the cluster center and the whole process is repeated starting with the calculation of the average position of data points closest to each new cluster center. Also the median can be similarly applied instead of the average in the calculation of the new position for the center point. It is often customary to repeat the clustering with several choices for starting points (de Hoon, et al., 2004) and to select the most useful result(s) using some clustering goodness measure (see chapter 2.3.4 for more information).

Many clustering methods assume that the clusters have spherical or elliptical shapes. K-means is especially a method that makes this assumption. In real life the shape of the cluster can be more arbitrary. Therefore also other ways of defining clusters have been tested and one option is based on the variations of the densities in the data space. It is natural to look for dense areas and define them as clusters and use the sparser areas in the data space as border areas for neighboring clusters. There have been some recent advances with these clustering methods (Getz, et al., 2000b, Sharan and Shamir, 2000). The possibility of having only the density as a limiting factor for the clusters is reasonable, but the drawback is the problems caused by overlapping clusters (Getz, et al., 2000a).
Other new statistical measures also enable the clustering without knowing the correct number of clusters. These are often based on Bayesian data modeling where the distribution of the data is modeled with density distributions and the number of clusters is allowed to vary (Medvedovic and Sivaganesan, 2002). Instead of giving one clustering solution these methods often create a large group of potential clustering results. These results tend to present a new analysis challenge in the combination of several clustering results (Medvedovic and Sivaganesan, 2002).

### 2.3.3.1. Self-Organizing Maps (SOM)

Self-Organizing Maps (SOM) (Kohonen, 1997) are one popular methods for clustering gene expression data (Hautaniemi, et al., 2003, Tamayo, et al., 1999). SOM is a grid of map units that functions similarly like the k-means clustering by looking for optimal placement of the model vectors associated with map units. As a result each map unit is a presentation for the data grouped inside the map unit, although some map units can also be empty. The main difference to k-means is that the SOM includes a grid structure that connects the map units to neighboring map units. The grid structure aims to ordered similarity presentation of the data where the most similar data points are being placed to the same map unit, similar data points are in neighboring map units and data points that are most dissimilar are placed wide apart on the map. The grid structure can be folded, shrank and/or stretched like an elastic sheet in the data space in order to obtain this.

Grouping of data points makes SOM similar to clustering algorithms whereas the ordered similarity presentation creates a visualization of the data similar to non-linear dimension reduction methods. The resulting SOM can be easily shown for example as a two-dimensional visualization where the obtained clusters can be directly used for analysis. Still it is often more recommended to use them as a preliminary clustering and to perform the final clustering using the SOM grid structure to join neighboring map units (Kaski, et al., 1998).

Grouping is based on connecting a model vector to each map unit and placing each data point to the map unit that has got the closest model vector (Best Matching Unit, BMU). BMU is chosen using the selected distance measure. Used model vectors have the same dimensionality as the data point vectors. When the data point is mapped to BMU the selected BMU and it’s neighboring model vectors are being moved towards the data point. During the teaching procedure the movement is big at the start, but it is slowly diminished during the iterative comparison of the data points and model vectors.

There are some parameters that have to be defined in the teaching process. The topology of the resulting SOM has to be defined. Usually the resulting SOM is a two-dimensional grid with either hexagonal or rectangular topology. In hexagonal topology (honeycomb structure) each map unit has six neighbors and in rectangular topology (chessboard structure) each map unit has four neighbors (see for example (Hsu, et al., 2003, Vesanto, et al., 2000)). From these two options the hexagonal grid structure is considered better as it has more connectivity between neighbors.
The size of the resulting SOM also has to be defined. A larger map will give a more detailed view but at the same time it can start folding too much and oscillating in the data space. A smaller map would be more robust including bigger clusters but it will be less able to model the finer details from the data space.

Another parameter that affects the behavior of the map in the data space is the stiffness of the resulting map. The stiffness parameter defines the size of the neighborhood that is being moved also every time the nearest map unit is being moved towards the analyzed data point. Usually the neighborhood function is a Gaussian distribution having two parameters: height at the center of the distribution, and radius. Here the center is the position of the closest map unit. Height defines how strong the movement of the closest map unit towards the data point is and radius defines how wide area will be similarly affected. Note that due to the shape of the Gaussian distribution the closest map unit will have the strongest movement and the further away the map unit is from the closest map unit the less it will move towards the data point. It should be noted that the optimal stiffness parameter depends on the size of the SOM map. The larger the map the bigger the neighborhood function should be.

Definition of these parameters usually requires background experience about the algorithm and the data and finding an optimal selection for the data usually requires repetitive testing of different selections of parameters. To make this selection more reasonable statistical models have been proposed for SOM (Lampinen and Kostiainen, 2002, Utsugi, 2000) enabling maximum likelihood search for the parameters.

**2.3.3.2. Tree-Structure SOM**

SOM has to be also initialized before starting the search and there are several ways for initialization. One could use other dimension reduction methods like Principal component analysis to generate starting dimensions for the SOM map or randomly selected dimensions (Vesanto, et al., 2000). One solution to initialization of the SOM is presented in tree-structured SOM (TS-SOM, (Koikkalainen, 1995)). Here whole data is first modeled by it’s average. Next a SOM structure of 2X2 grid is fitted to the data. This will present a very crude partition of the data that is not usually by itself useful. Still it can be used to guide the next steps as each map unit in the grid is used as an approximate starting point for four new map units in 2X2 grid on the next level of the SOM, resulting into a 4X4 map. SOM is enlarged this way by iteratively fitting the 2X2 structure to each previous level map unit until the needed size of SOM map is reached. Also the search of closest map units is limited in the TS-SOM to map units that are approximately in the same area as the earlier map nodes that included the data point. This method was reported to result in a considerable speedup in the SOM teaching process (Koikkalainen, 1995). Notice that the method results in a possibility to view data with several SOM grid sizes, although the grid sizes are limited to 2X2, 4X4, 8X8, 16X16 etc.
2.3.3.3. Methods similar to SOM

Besides the enthusiastic response in the literature to the SOM algorithm, there has also been some criticism. One cause of criticism is the requirement of fixed grid structure before the analysis. There are many methods that have been developed to overcome this requirement such as Growing Network Structures (Fritzke, 1994) and Grow When Required networks (Marsland, et al., 2002). One of the methods, Growing Self-organizing Maps, has been tested with gene expression data analysis (Luo, et al., 2004). A drawback with these methods is that they usually require a definition of one or few additive parameters that guide the growing process for defining new map units and also possible deletion process of map units with little importance (usually neurons with very few or no members at all).

2.3.3.4. Visualization of the SOM result

Although the SOM already presents a useful two-dimensional visualization it does not show all the details from the data. One detail that is not shown in the standard SOM visualization is the difference between the neighboring model vectors as they can vary a lot. Two well-known methods were used in this work to visualize these differences. One is the U-matrix method (Ultsch and Siemon, 1990) where a colorplot is added on top of the SOM map to visualize these differences. This usually shows clusters as a bright areas and the in between cluster areas with dark in the SOM map. Another way is to use other dimension reduction methods to visualize the similarities of neighboring map units (for example Sammon’s mapping (Sammon, 1969)). Usage of SOM before the visualization benefits the visualization by lessening the calculation burden in the visualization process.

Contribution of different dimensions can also be visualized on top of SOM map (Vesanto, et al., 2000). When the number of SOM map units is small and when the dimensionality of the data is small the map unit model vectors are usually visualized inside the SOM map unit (see for example (Tamayo, et al., 1999)). As the number of map units and/or the number of dimensions in the data vector increases it becomes increasingly harder to create an understandable presentation with the earlier visualization method. Then it is better to visualize values for one dimension in the SOM map unit model vectors on top of the SOM map one at a time (a component plot, (Vesanto, et al., 2000)). Both these methods enable the analysis of contribution of different dimensions to the grouping of the data at a selected area of the SOM map.

2.3.4. Evaluation of the clustering results

One drawback with clustering is that often the user has to define a parameter or parameters that effect strongly on the clustering results. A parameter could be for example: selected number for clusters (k-means, cutting a hierarchical tree with selected number of clusters), maximum size or maximum variance for the clusters (for example method in (Heyer, et al., 1999), or cutting a cluster tree with maximum size or variance)
or maximum or approximate distance between two neighboring data points (Ben-Dor, et al., 1999, Getz, et al., 2000b, Sharan and Shamir, 2000). The selection of optimal parameters usually requires considerable testing and evaluation of the clustering results. As all the possible parameter options will usually produce clusters, and also different clustering methods can similarly produce different results, the important question is how to select the important results among the several possible options. Two types of testing methods exist for the testing of obtained clusters: external and internal indices (Jain AK, 1988)

Internal indices use only the information in the clustered expression data to evaluate the obtained clusters. With internal indices the evaluation can be based on the testing of obtained clustering with a part of the data that was left out of the analysis or it can be based on the comparison of the distances between the data points inside the clusters when compared to the distances of data points between the data points in different clusters (see for example: (Dudoit and Fridlyand, 2002, Tibshirani, et al., 2001, Tibshirani, et al., 2000)). These are methods that do not require any other information for the evaluation of the obtained clustering results but the data that was given for the cluster process.

External indices use some other data to evaluate the resulting clustering by testing how much the obtained clustering result correlates with other data. This other data should be selected so that it highlights such results that are considered important in the obtained clustering solution. As a simple example, groups of genes that are known to behave in a similar way in the experiment setup. On the other hand, a very poor example of the evaluation information, would be the localization of the measured spots on a microarray, as this would emphasize such clustering solutions that enhance local artifacts in the data. The latter type of information could still be used as negative control data to make sure that unwanted signal has not created the cluster structure in data. Evaluation using external indices is similar to evaluation of preprocessing steps (see end of the chapter 2.2.3.4.) and the two can be actually combined (Gibbons and Roth, 2002).

2.3.5. External information for the evaluation of the clusters

Normally verified information of similarly behaving genes in the experiment setup is not available. Therefore other genomic data sets have to be used to evaluate the obtained results. The use of other data sets should be based on the reasonable assumption that the two data sets should correlate and have different error sources. As an example of suitable data sets, it has been shown that the clusters obtained with protein-protein interaction data and gene expression data correlate (Ge, et al., 2001) or that the clusters obtained from sequence clustering correlate with the ones obtained from gene expression data (Bilu and Linial, 2002).

One natural and common choice of information for the evaluation of results would be the information about the functionality of the genes, as it is natural to assume that if a certain function needs regulating in the test setup then several genes from the same functional group would show it (see for example (Eisen, et al., 1998, Gibbons and Roth, 2002, Oja, et al., 2002)). Annotated genes can be classified to categories using several possible
features as a starting point for the categorization. Genes have been classified, for example, according to their cellular components, protein complexes and cellular functions. Often these functional annotations are presented in hierarchical fashion showing different level of detail in different classifications. As an example genes can be classified to Metabolism, Amino-acid metabolism and to Biosynthesis of amino acids. Currently the Gene Ontology (GO) project has aimed at standardization of such classifications (Ashburner, et al., 2000, Ashburner and Lewis, 2002). Later chapters (4.4) demonstrate the usage of two databases, Saccaromyces Genome Database (SGD, (Weng, et al., 2003)) and Munich Information center for Protein Sequences (MIPS, (Mewes, et al., 2000)) for evaluation purposes.

3. Aims of the study

The presented work had several levels with different aims. The first level includes the preprocessing of gene expression data. Here the aim was to look for measures that can reliably evaluate the change in the expression despite the error signal in the expression measurements (publications I and IV).

The second level included the analysis of gene expression data with clustering. At this level our aims were to: test the usefulness of Self-Organizing Maps to clustering and visualization of gene expression data (publications II and III); develop methods to extract important features from the gene expression SOMs (publications III and IV); develop methods to compare and evaluate important features found from the gene expression SOMs (publication IV). These projects proposed the usefulness of a method to search directly the clusters presenting interesting correlation with some gene functional classes in the data (publication V).

The last level was the analysis of the obtained gene groups. At this level, the aim was to define methods for intuitive analysis of obtained clusters (publications IV and V). Methods presented at this level are partly unpublished.

4. Materials and methods

4.1. Data sets

4.1.1. Publicly available data sets

Two types of data sets were used during the analysis. One group is the publicly available data sets. These are usually large including from seven to 300 measurements (dimensions) for each gene. A large number of dimensions allows a wide variety of distance measures. All the public data sets used here were done analyzing a baker’s yeast (Saccharomyces cerevisiae) with drastic treatments causing strong changes in the gene
expression. Yeast was one of the first organisms that was analyzed with cDNA chips that included all the estimated Open Reading Frames (ORFs) for the analyzed organism. Also a simpler organism creates a good starting point for the analysis methods before moving to more complex organisms.

The first data set (DeRisi, et al., 1997) used included the measurements for over 6000 yeast genes while shifting from fermentation (energy production without oxygen) to respiration (energy production using oxygen). The data set shows the changes occurring when the energy source is switched and also the stress and adaptations caused by less optimal growth conditions. Data set will be called DeRisi data set in the later analysis steps.

The second analyzed data set (Eisen, et al., 1998) monitors yeast in various types of stress situations and in a cell cycle (table 1). Stress treatments show yeasts response when nitrogen or a carbon source is poorly available (sporulation and diauxic shift), when the temperature is not optimal (cold and heat shocks), or when chemical compound disturbs the protein synthesis (reducing environment, DTT). Cell cycle data sets on the other hand monitor the correlation of gene expression to cell cycle. Three different methods were used for synchronizing the sample cells to same cell cycle phase producing three data sets. Only a part of yeast genes is included in this data set (2467 genes). This data set will be called Eisen data set in the later analysis.

Table 1 – experimental conditions in the Eisen data set

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Number of time points</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell-cycle: Alpha factor arrest and release</td>
<td>18</td>
</tr>
<tr>
<td>Cell-cycle: Elutriation</td>
<td>14</td>
</tr>
<tr>
<td>Cell-cycle: cdc15 arrest and release</td>
<td>15</td>
</tr>
<tr>
<td>Sporulation</td>
<td>11</td>
</tr>
<tr>
<td>Heat Shock 25°C-37°C</td>
<td>6</td>
</tr>
<tr>
<td>DTT Shock 1mM</td>
<td>4</td>
</tr>
<tr>
<td>Cold Shock 25°C -11°C</td>
<td>4</td>
</tr>
<tr>
<td>Diauxic Shift</td>
<td>7</td>
</tr>
</tbody>
</table>

The table presents the experiments done in the Eisen data set (Eisen, et al., 1998) and also shows the included time points. A detailed table of experiment conditions can be found from internet (http://www.pnas.org/cgi/data/95/25/14863/DC1/1).

The third data set (Hughes, et al., 2000) is also the most complex data set. It monitors the response of the yeast at the gene expression level to the lack of different genes. This is analyzed with different mutated yeast strains each lacking a selected gene or with different bioactive compounds that block the function of a specific gene. Altogether the functions of 300 yeast genes with various functionalities were separately blocked. Each of these created various stresses to yeast cells causing cells to compensate the missing functionality of the inactive gene with parallel pathways and other means. The data set will be called Hughes data set in the later analysis.
The Hughes data set includes also a large analysis of the variance of the obtained results when control samples of normal yeast cells were compared against each other. This data shows the variance that one could expect from similar samples and it was used to develop a mathematical model for error in gene expression data (Hughes, et al., 2000). This control data also gave its own approximation of the error. In the original work, this approximation replaced the error model whenever the error estimated from the control data was bigger than the calculated error. As the actual test data set includes too few repetitions and therefore has a big risk of obtaining false positives, the more pessimistic error was selected similarly here.

There is also another control data set hidden in the main Hughes data set that has not been used by other authors. These are genes for example from other organisms or measurements for spot areas with no attached gene. This data presents the measurements that one could get with a gene that does not exist in the sample when the earlier control data set shows the measurements that one obtains when the expressed genes from similar samples are being measured. The data also included Non-annotated ORF sequences. These were the ORFs that were not annotated in the first analyses of the yeast genome. Control data sets were used to evaluate the obtained signals in the actual data set to see if they should be considered significant or a result of the background noise (publication V).

4.1.2. In house produced data sets

In house produced data sets are obtained from the analysis of effects of anti-psychotic drugs on rat brains. These data sets are smaller than the public data sets usually including 2 to three different treatments for each gene. Also the treatments in these data sets are less drastic than those with yeast resulting in weaker responses in gene expression. Weaker signals are harder to detect among the random noise in the data.

Two in house data sets were analyzed. The first one included measurements from prefrontal cortex region of rat brains with two different treatments of the model drug clozapine. Treatments were acute and chronic. In acute treatment three time points (1, 6 and 24 hours) after treatment were measured. In chronic treatment animals were first injected with haloperidol or clozapine for 17 days. Samples in chronic treatment were taken after washout period of 2 hours or 1 or 6 days. Washout period is the time during which the animal is not treated with drug. All of these measurements were done using cDNA macro arrays. A reference data set was animals similarly treated with saline (see publication IV).

The second data set included measurements from the rat brain parietal cortex tissue after dizocilpine maleate (MK-801) treatment. MK-801 is a compound that creates psychotic effects. Two MK-801 dosage sizes, 1 and 5 mg/kg, were selected for the analysis of gene expression. Here the reference data set was also animals treated similarly with saline (see publication I).
4.2. Preprocessing

4.2.1. Preprocessing of public data sets

Preprocessing of the public data sets was usually done similarly to work done in original articles describing the first analysis of each data set although some differences have been presented. The aim of minimizing the differences in preprocessing steps is to make the obtained results comparable with the original analysis. The DeRisi data set (DeRisi, et al., 1997) was analyzed using log-ratio measurements similarly to the original article. Euclidean metric was selected as a distance measure for clustering. Use of Euclidean distance simplified the analysis, as there was no need to filter the genes with no significant variance in their expression profile (see chapter 2.2.3.4). Visualization of the results actually showed that the genes with no regulation created their own group (publication II).

The Eisen data set (Eisen, et al., 1998) was otherwise preprocessed similarly as the DeRisi data set but the distance measure was different. Here a distance measure similar to Pearson correlation was used instead of the Euclidean metric. Similarity to Pearson correlation was obtained by scaling the average of data lines to zero and variance to one and using the dot product as a distance measure. Differences between distance measures have been discussed more in detail in the literature review (see chapter 2.2.3.4).

It should be noted that the genes with too little variance in their expression profile should be discarded when using Pearson correlation. This was not done when analyzing Eisen data set as it was estimated in the original analysis that all the genes included to the analysis had significant variance somewhere in the expression profile (Eisen, et al., 1998). Significant variance of all the genes is probably a byproduct of the selection where only the genes that were functionally known at the time of the publication were taken to the Eisen data set. Missing values in this data set were replaced with a nearest neighbor type estimate implemented in the Visual Data software (Visipoint Inc. Kuopio, Finland). This method is similar to one presented in the literature (Troyanskaya, et al., 2001)

The Hughes data set (Hughes, et al., 2000) was preprocessed differently from the other public data sets. The major factor causing different preprocessing was that the Hughes data set included estimates for the error of log(ratio) values. The error estimate was used in the original analysis to filter away the effects of additive and multiplicative error. Here the error estimate was used to scale the obtained log-ratio measurements and Pearson correlation of the scaled values was used as a distance measure. The combination of the Pearson correlation and error scaled measurements was also used in the original article (Hughes, et al., 2000) and it has been shown to be one of the optimal distance measures for the Hughes data set (Oja, et al., 2002). The bigger one of two error estimates for each gene was always selected as was explained earlier.

Some of the logs of ratio-values were reported as missing values (or Not a Number, NaN) in the analysis of Hughes data set. The measurement was reported as a missing value if
the log of ratio value was missing or was reported as NaN, if the error for log of ratio value was approaching infinity or was reported as NaN and/or if the p-value for log of ratio value from the background noise distribution was 1. In all above cases there was a strong reason to believe that the measurement does not have any significant signal and therefore it should be omitted from the analysis. These omitted values were replaced with zero. Replacement with zero is based on the assumption that the missing values are coming from log(ratio) values with very large error causing the error divided value to approach zero (publication V). Replaced values should have a negligible effect on the distance measure as the average of the data lines was observed to be close to zero and the closer the value is to average the smaller is its effect on correlation measure. This issue is discussed more in the original article (publication V).

4.2.1.1. Novel method for filtering low variance genes

It was earlier mentioned that the genes with too low variance have to be filtered when Pearson correlation is used as a distance measure (see chapter 2.2.3.4). Here a novel method for the filtering step was developed for the analysis of Hughes data set (publication V). A measure for evaluating amount of signal was first selected. Here the variance coming from the changed expression was compared to the total variance, which included the variance of changing expression, and the variance caused by measurement error (publication V, see the supplementary text). The variance in the data line could have been used similarly. This measure is called Ratio of Variance (RoV) similar to original article.
Figure 4. Comparison of signal vs. variance with two control data sets and test data set.

The figure shows the distribution of the obtained results from the analysis of the amount of variance caused by signal when compared to the total variance for each gene. Line marked with C shows the resulting distribution for the mutation data set and the lines marked with A and B show the same for two control data sets. B indicates the results from the comparison of the control samples and the A indicates the results with genes not present in the samples. The vertical line shows the selected cutoff. Notice that the amount of signal (shown on the X-axis) is clearly larger in the mutation data set. The histogram distribution for B and C data sets is being visualized with 200 bins but the A is only with 10 bins as it includes significantly less members (~1/100 of the other data sets). A similar figure is originally published in publication V.

The obtained measurements were used to see if the data lines from the experimental data set and from control data sets could be separated with the selected measure. This was performed by plotting the distributions of the values from control data sets and from test data set (figure 4). Next a cutoff value was selected to separate the actual data set from control distributions. As a result the cutoff should separate the genes with variance only caused by noise and genes having significant variance in their expression profile. The borderline between two groups was selected manually by taking the larger one of the 95th percentiles as a cutoff (publication V). The selected cutoff discarded a small portion of the data set (980 genes out of 6357) especially when compared to the original article cutoffs but still it can be seen from the histograms of the test and control data sets (shown in figure 4) that the cutoff is quite conservatively selected. Also the obtained clustering
results pointed to strong correlation between the functional classes and clusters increasing the reliability of the obtained data (see table 2).

### 4.2.2. Preprocessing of in house data sets

Measurement errors caused more problems with the analysis of the in house data sets. The most increased and most decreased ratios were observed to point to array spot areas having practically no signal at all in either the test or control image. This was observed when analyzing the in house data sets with simply the expression ratios between sample and control and confirming the result with visual inspection of the spots. Obtained results can be explained by the additive errors in the data sets. The smaller the original measurements the stronger the effects of additive errors are. If for example a measurement stays at a constant level of 4 an additive error of size 2 is expected to create ratios of size 3 (= 6/2) and 1/3 (= 2/6). On the other hand, if the measurement stays at the constant level of 100 previous additive errors would have only a very small effect on the resulting ratio.

To correct this phenomenon we used a method where the standard equation for ratio was corrected by adding a constant to the divider and to the dividend. This constant was defined to be proportional to the average of the measurement group in question. The added constant forces the resulting ratio closer to 1 when values used to calculate the ratio are small. Use of this method requires that the average or the median values of the two data sets have been scaled to the same value. If this is not done, the ratios for unchanged genes at the higher expression levels and at the lower expression levels will show different results.

The size of the correcting constant has an effect on the results. If it is too small the correction is not strong enough. If it is too big, the correction is too strong and discards truly changed measurements from the analysis. Suitable size for the constant was estimated from ratio vs. average scatter plots by trying to select the smallest possible value for the constant that would provide a reasonable correction (see figure 2 for example). The aim was to modify the distribution of the ratios so that the truly regulated genes would be selected to analysis instead of the genes with low-intensity measurements showing mostly the error signal in the data. Usually we tested the constant of the size 0.1*average to 0.4*average. A constant of size 0.2*average of measurements was selected to the analysis and it can be seen from the figure 2 C that it lessens the effects of the exaggerated variance among the low intensity measurements. Also the visual analysis of the chip areas confirmed that the selected spots really showed change in the expression (data not shown). Further processing of these data sets was standard. The log of result from corrected ratio was taken to further analysis and Euclidean distance was taken as the distance measure.

### 4.3. Algorithms
4.3.1. Clustering algorithms

4.3.1.1. Self-Organizing Maps

Two different Self-Organizing Maps (SOM) algorithms were applied here. The DeRisi data set was clustered with Tree Structure SOM algorithm implemented in Windows data analysis software (Visual Data, Visipoint Inc., Kuopio, Finland, www.visipoint.fi). The Eisen data set was clustered using SOM-PAK software (http://www.cis.hut.fi/research/som_lvq_pak.shtml). Further analysis of the Eisen data set was done in Matlab environment (Mathworks, Natick) using Matlab SOM toolbox (http://www.cis.hut.fi/projects/somtoolbox/) and specially tailored functions.

4.3.1.2. Visualization and clustering of SOM results

To obtain more information from the SOM presentation of the data, we visualized the distances between map units using U-matrix method (publication III) and Sammon’s mappings (publication II). These methods enable the search of the possible cluster regions from the SOM presentation by showing the differences between more and less dense areas found by SOM.

Although most of the gene expression data analysis tends to concentrate on analysis of cluster structure from the gene expression data, we have also found it interesting to look for functional classes and other classes for analyzed genes that show localization in the resulting SOM visualization (publication III). We analyzed these areas showing such localization from SOM visualization as potentially interesting areas for the analysis.

One of the key questions while analyzing cluster areas is the question of the main differences in the expression values between the selected area and the rest of the data. These were visualized using component plots from SOM or presenting the expression profiles for the genes in the selected cluster (see chapter 2.3.3.4). Component plots present the visualization of the values for one dimension in map unit over the SOM map with a color gradient.

4.3.1.3. Hierarchical Clustering

Hierarchical clustering was used in the analysis of the Hughes data set as a preprocessing step for the cluster selection method. Agglomerative clustering was used where data points and/or clusters of data points are being joined repetitively until only one cluster remains. Analysis was performed using three different clustering methods: average linkage, complete linkage and Ward’s minimum variance method. Hierarchical clustering was performed in the Matlab environment using modified Matlab built in clustering functions. Modified functions enabled a more optimal sorting of the resulting cluster tree nodes making the resulting visualization more informative (publication V).
4.4. Gene annotations and classifications

Analysis of gene clusters obtained from the data set is not typically easy. It can occasionally be hard and time-consuming to notice the common features for the grouped genes. In order to ease the analysis of the obtained gene groups, and to enable statistical evaluation of the clusters, we have used functional and other classifications of genes. Classifications can be used to highlight the important features associated with the cluster in question. Functional categorizations used in this work for yeast were obtained from SGD (Weng, et al., 2003) and MIPS databases (Mewes, et al., 2000). Both of these were used with Hughes data set (publication V) whereas only MIPS was used with the Eisen data set (publication III). Although these functional databases included hierarchical classifications of genes, no specific level from the hierarchy was selected for the functional annotation here. Instead, all the levels in the resulting hierarchical classification were used for the annotation and categorization of the genes as it was assumed that the data could include clusters with various functional details.

The Eisen data set also included short functional descriptions about the genes. These were used as a starting point for keyword searches where occurrences and co-occurrences of keywords associated to gene functions were looked for (publication III). Occurrences of keywords were used in a similar way like the functional classes to enhance the analysis. Keyword searches usually enabled analysis of more detailed groups than the original functional classes. One of the articles also included in house gene classifications (publication IV).

4.4.1. Correlation measure

Despite the use of functional classes it is hard to evaluate the significance of the observed grouping. As an example, it would not be a surprise to find out that 2/3 of the cluster includes cytoplasmic proteins when the whole data set includes the same proportion of cytoplasmic proteins. Still it can be easy to create erroneous conclusions about the observed proportion of cytoplasmic proteins, although the same result could easily come by simply taking randomly a similar sample from the data set. This is especially dangerous when analyzing custom tailored gene chips including genes associated to certain biological phenomena and therefore having a very skewed selection of gene functions.

A natural way to analyze importance of different functional groups would be to use ratio of number of members of the functional class in the cluster vs. the size of the cluster. This is often referred to as “specificity” in the classification literature. Also, the number of members of the functional class in the cluster can be compared to the size of the class. This is often referred to as “sensitivity” in the classification literature. Specificity and sensitivity have defects as measures. Specificity does not take into account how common the functional class is. Therefore it would give an optimal result for a cluster of one gene
that belongs to a class that most of the genes belong to. Sensitivity does not take into consideration how many genes there are in the cluster that do not belong to the functional class. So sensitivity would give an optimal result with a cluster that most of the genes belong to when it includes the only gene that is a member of the analyzed class. Both of the previous cases are actually quite likely to occur just by accident. Also, neither of these measures actually answers the most interesting question: what is the probability of the similar number of class members in a gene group by chance.

Sampling without replacement is a standard statistics method that answers the previous question. It analyzes how many ways there are to obtain the observed result from all the possible ways of how sample could be taken from data set and gives the probability of the observed number of class members by random \( \left( A_k = \text{“cluster contains } k \text{ members of gene class“} \right) \):

\[
P(A_k) = \frac{{K \choose k} {N-K \choose n-k}}{{N \choose n}} \quad \text{(eq. 1)}
\]

Here \( K \) is the size of the cluster, \( k \) is the number of class members in the cluster, \( N \) is the size of the entire data set and \( n \) is the number of class members in the whole data set. All the possible outcomes from this sampling create a hypergeometric distribution. Sampling without replacement actually functions here as a zero model (M0) defining the probability that the observed sample of data has no significant difference to a random sample from the same data. Zero models are regularly used in statistics to evaluate whether it is reasonable to make any conclusions at all from the observed data or is it just a random sample.

The drawback with sampling without replacement is that as the size of the sample increases the number of possible outcomes usually increases at the same time. With the increasing number of possible outcomes, the probability will also be distributed among bigger number of outcomes. As a result, sampling without replacement reports smaller probability for cases where there are more possible outcomes. To correct for this kind of phenomena there are usually two methods used: first, the comparison of model describing a data with a significant signal (M1) to M0 using Bayes factor; second, the calculation of p-value over the tail area that forms the group of more extreme results than the one observed using hypergeometric distribution. The latter has been used here and it can be presented with the equation:
Here $E(X)$ represents the expected value (mean) of the distribution that is in this case a hypergeometric distribution with parameters $N,K$ and $n$. $x_{\min}$ and $x_{\max}$ represent the minimum (bigger of two values 0 and $n-N+M$) and maximum (smaller of two values $K$ and $n$) values that $x$ can get in the hypergeometric distribution. $A_{x \geq k}$ means “cluster contains $k$ or more members of gene class” and $A_{x \leq k}$ means “cluster contains $k$ or less members of gene class”. If the obtained result is bigger than the mean of distribution the probabilities for possible results larger than the obtained one will be summed to p-value. Similarly if the result is smaller than the mean of distribution the probabilities for smaller possible results will be summed to the p-value. So the tail-area of the hypergeometric distribution is always summed here to p-value. Notice that although the probability values close to the average value of the hypergeometric distribution originally differ depending on the number of possible outcomes as a result of summing over the tail area the p-value would be close to 0.5. The obtained measure is actually Fisher’s exact test (Agresti, 1992, Fisher, 1922).

The measure has still some drawbacks for our purposes. First of all, the normal scale leaves most of the scale for insignificant values like values between 0.1 and 1 while at the same time values close to zero are given very little scale. It is natural to consider that the differences between probabilities close to zero (see table 2) are more important than say the change from 0.1 to 1. Therefore the negative log-base 10 of the p-value is reported instead of the original p-value. The logarithm gives more scale to the p-values approaching zero. The measure still includes one unwanted feature. It treats significant positive correlation (functional class is over-represented or enriched in the sample) and significant negative correlation (functional class is under-represented) similarly although they present totally opposite cases. Defining a changing sign for the log of the p-value solved this. The sign of the log-p-value was defined positive when the correlation is positive and negative when the correlation is negative.
The obtained measure can be used to divide functional classes into three groups: significantly enriched functional classes, significantly under-represented classes, and functional classes that do not correlate significantly with the analyzed group. Significant results can be selected by taking $-2$ and $+2$ as limits (when using 0.01 as a limit). The measure can be also used to rank functional classes or clusters as is described below. The correlation measure is very similar to that used in the literature (Wu, et al., 2002). Two measures are identical when the functional class is being enriched but the difference between the two measures is that the correlation measure presented here can also report significant negative correlation (see figure 5).
4.4.2. Applications of the correlation measure

Two applications for the presented correlation measure have been tested in the current work (IV and V). The first uses manual and the second automatic selection of clusters. Cluster areas were obtained in manual selection by taking local cluster areas in the SOM map presentation and using the enrichment of interesting gene classes in the obtained clusters to choose clusters for further analysis. Cluster areas were selected on the basis of observed localization of analyzed functional classes and/or on the basis of interesting

Figure 5. Comparison of two correlation measures.

The two profiles show differences and similarities between log-p-value measure used here and the -log of p-value measure used in literature (Wu, et al., 2002). The figure presents the results for each possible outcome for a cluster of 50 genes from the pool of 200 genes with half of them belonging to the monitored class. This corresponds to hypergeometric distribution with parameters $N = 200$, $K = 50$ and $n = 100$ (see text for details). The X-axis shows the number of genes in the cluster that were observed to belong to the monitored class and the Y-axis shows the obtained results (-log-p-value for Wu et al’s method and equation 3.). Note that equation 3 also highlights situations when the negative correlation is significant and one could select values such as +/- 2 (equal to p-value <= 0.01) as the borderlines for significant enrichment and under-representation from the vertical axis. The vertical line presents the mean of the associated hypergeometric distribution and the dashed vertical lines present the mean +/- standard deviations. The figure is originally presented in the supplementary text of the publication V.
local expression profile. This method actually presents the next step in the analysis of the class distributions on top of SOM clustering done in publication III as it enables the selection of the areas from the SOM that show strongest enrichment of functional classes of interest and it also enables the comparison of overlapping clusters. Automatic selection was used with hierarchical clustering methods and it is explained in the later chapters (chapter 5.2).

Another level of application was the analysis of obtained clusters where the evaluation of the clusters was aided with a list of most significantly enriched functional classes and most significantly under-represented functional classes in the cluster. Functional classes were sorted according to their p-values and listed starting with the class that presented the most significant correlation. Besides the reported functional class and its log-p-value, the results included also observed number of class members in the cluster, total number of class members in the whole data set, the size of the cluster, and expected number of class members in the cluster (average value from hypergeometric distribution). The additional information was used to confirm the results. An example of the output is shown in table 3.

4.4.3. Searching method for clusters enriching functional classes for cluster trees

Work done with SOM required user intervention. Cluster areas had to be selected by going through distributions of several classes on top of SOM and the results had to be manually evaluated. Therefore, an automatic method was considered in publication V. A simple way to do similar search of interesting cluster areas from the data set was to move the analysis to hierarchical clustering and to look for clusters presenting strongest enrichment. The idea for enhancing the analysis of the cluster tree is that at some point in each branch of the cluster tree there is a cluster that presents an optimal enrichment of the functional gene class that is most over-represented in that part of the cluster tree. The previously presented correlation score for the most enriched monitored gene class in each cluster was used to measure the enrichment. The cluster showing the strongest correlation was always selected. After selecting a cluster, the tree branch where the cluster was found was not considered anymore important for obtaining correlation with the reported gene class. Therefore, the clusters in the same cluster tree branch showing the strongest enrichment for the same functional class as the cluster reported are excluded from the search. This step is called silencing and its principle is presented in the figure 6. The search procedure and exclusion are repeated again until some stopping criterion is met. In the presented work, the minimum limit for log-p-value for reported clusters was chosen.
Figure 6. Principle for finding clusters with optimal correlation with the analyzed functional classes.

In fig. 6 A, the cluster tree and the used log-p-value correlation measures for two imaginary functional classes (black and grey) are being obtained for the analysis.

The strongest enrichment (bigger value) is stored for each cluster for the step in fig. 6 B, where the method looks for the best score in the whole tree (cluster F with score 8 for black class). The cluster is marked as Best Scoring Cluster (BSC).

In fig. 6 C, the clusters above and below in the hierarchy are being discarded from the analysis if they have the same functional class with strongest enrichment as the previously reported BSC. This step is called silencing and it discards the repetitive hits to the same cluster tree branch with the same functional
class. Notice that the small cluster (cluster A) enriching the grey class is not silenced. The figure is a modified version from that originally published in publication V and a color version of it can be seen at the publisher’s web site.

5. Results

5.1. Screening of genes from in house data sets

Work presented in publication I consisted mainly of screening for regulated genes. Measurements from control and sample macro array were normalized using the house keeping gene (β-actin). The analysis concentrated on genes assumed to be regulated in the MK-801 response. Observed regulations were small (often less than 2) so the reliability of these regulations was weak. Therefore some of these regulations were confirmed with in-situ hybridization (see publication I).

5.2. SOM clustering

5.2.1. DeRisi data

The DeRisi data (DeRisi, et al., 1997) was first used to teach the SOM map which in turn was used as a starting point for the visualization with Sammon’s mapping. Resulting SOM and Sammon’s mapping visualizations (figure 2 in publication II) showed one principal axis of differential regulation. At one end of this axis the genes are down-regulated at the end of the diauxic shift (area close to C3 in figure 2 in II) and at the other end the genes are up-regulated in diauxic shift (area close to C1 in figure 2 in II). This gradient of the response forms the main axis although some other differential responses can be found. For example, close to the same area with the genes having their maximum up regulation at the last time point are also the genes having their maximum up-regulation in the second to last time point.

There were actually four groups of genes that were reported in the original article (DeRisi, et al., 1997). These four groups were the genes that had strongest up-regulation in the last time point, genes that had strongest up-regulation in the second to last time point and opposite groups that had strongest down-regulation in the last time point and the group that had strongest down-regulation in the second to last time point. All these four groups were localized in the SOM map (data not shown). We took sample map units from the three regions and showed that the genes clustered together were very similar in the functional level. The cluster with strongest up-regulation in the last time point had mainly genes from the oxidative energy production. The cluster with strongest up-regulation in the second to last time point included stress related genes like heat shock proteins. The third cluster showing strongest down-regulation in the last time point included almost purely ribosomal proteins. Down-regulation of the ribosomal proteins is a typical reaction of the yeast cell to weakening of the growth conditions and
environmental stress. It was also observed that neighboring clusters at these regions had genes with similar functions. The areas at the middle of the SOM map were less coherent in function and presented mainly random distribution of the gene functions.

### 5.2.2. Eisen data set

Previous analysis of the DeRisi data concentrated on single SOM map units. It should be noted that usually the local map areas, including several map unit, form single clusters. Therefore the analysis of Eisen data (Eisen, et al., 1998) was tested with a larger SOM map putting the emphasis of the analysis more on the local SOM map regions.

The obtained SOM visualization was further enhanced mainly with two methods: visualization using U-matrix, and visualization of many gene classes on top of SOM map. Usually the cluster structure created by the U-matrix (see figure 1 in publication III) was compared with distributions of gene classes (see figure 2 in publication III). Gene class distributions enabled the functional annotation of the obtained clusters (shown in figure 1 in publication III). It should also be noted that some of the areas that were associated with gene classes are actually in between clusters, like the darker area V in figure 1 (publication III). This exemplifies that the analysis should probably not be limited to view just the denser regions in the data space like in classical cluster analysis, since areas in between the clusters can possess important features.

Analysis of the distribution of the gene classes showed that some of them were divided into several local groupings in the SOM visualization instead of one group. This raises a question of possible functional heterogeneity of the functional class in question. One such group that was further analyzed was ribosomal proteins. It was already shown earlier in the original analysis of the data (Eisen, et al., 1998), that cytoplasmic and mitochondrial ribosomal proteins are quite differently regulated, for example, in diauxic shift data (DeRisi, et al., 1997). This information was confirmed in our analysis, as we were also able to obtain separate areas for mitochondrial ribosomal proteins and cytoplasmic ribosomal proteins. In addition, our visualization showed an outlier group of mitochondrial ribosomal proteins. These showed different regulation from the rest of the mitochondrial ribosomal proteins mainly in cdc15 arrest and release treatment. Also, while analyzing cytoplasmic ribosomal proteins we noticed that they tend to be localized differently in the SOM map than the ribosomal structural RNA genes. This is surprising, as one would expect the protein and RNA members of the same complex to be regulated in a similar way.

Besides the clusters and SOM map areas with clear functional associations, there were map areas where the connection to some functional class or classes was not clear-cut. These were analyzed by looking at the expression profiles for the genes from the selected area (like in figures 4 and 9 in publication III). Also, the difference in the expression profiles between the cluster area and the neighboring regions to the cluster area can be analyzed. One cluster area analyzed in this way was the region in the upper left region (figure 5 in publication III). The analysis showed that the genes from this cluster area are
mostly upregulated in sporulation. In a more detailed analysis this cluster was divided into 8 subclusters (4X4 som map units). This revealed that different parts of the cluster show upregulation at different time points so that the upregulation starts first in the subclusters in the right part of the SOM cluster and travels through the cluster and upregulation occurs last in the clusters in the left part (data not shown). Similar findings where reported by Chu and coworkers in the original analysis of the sporulation data (Chu, et al., 1998). Pooling the data into a single cluster like in k-means clustering would make it harder to observe these details. A similar kind of area was in the left edge (see figure 6 in publication III). This area showed upregulation of genes mainly in some time points of sporulation, heatshock, and DTT treatment. This cluster area was later observed to include genes associated with protein folding and stress response confirming that it was a stress response cluster.

Besides analysis of the functional classes one at the time, we also viewed if there were interesting co-localizations. We noted an unexpected grouping of genes associated with carbohydrate metabolism at the area close to histone proteins (publication III). We used a keyword search (protein glycosylation) to analyze the distribution in more detail. Keyword search showed that protein glycosylation genes are slightly clustered at a region close to histone proteins (see figure 9 in publication III). We also viewed the expression profiles for the histone proteins and glycosylation proteins at the area to confirm the similar expression. Profiles showed similarities in cell cycles and in sporulation. This proposed an unknown association between these functions (publication III).

5.2.3. Selection and evaluation method for clusters from SOM

As previous results show the analysis applying functional classes has benefits but often one problem remains: When observing the area of localization of functional class how can one select a region for further analysis? Clusters on top of SOM can be made larger or smaller by joining or excluding border area neurons to the cluster. One way would be to concentrate on the obtained expression profiles and to base clustering to U-matrix or Sammon’s mapping, but as the previous results showed, the areas that are weak clusters or lie in between bigger clusters can be interesting. Therefore the enrichment of the interesting gene classes was selected as a measure for optimal cluster size for the further analysis steps.

This method was used to analyze several potentially interesting clusters (publication IV). The aim was to view gene expression data presented on the SOM level to find the regions where some of the analyzed gene classes show strong grouping. This approach does not exclude the normal SOM analysis of the cluster regions and the expression profiles (as shown in publication III), but rather complements it. The analysis method was applied here using specially designed gene classes instead of the standard functional classifications. As a result, the search found two presynaptic function based clusters. Note that finding and evaluating these local areas with enriched functional gene classes could be difficult without these methods (publication IV).
5.3. Results from cluster selection method for hierarchical clustering

Results with SOM showed the usefulness of using gene classes in evaluation and selection of the cluster areas (publication IV). Still, there is a lot of manual work included making the method dependent on the analysts input. Manual work makes it also hard to evaluate the results against the results obtained from randomization. In randomization, the method is tested repetitively with randomized connections between the genes in the class table and genes in the expression data. Randomization can evaluate the log-p-values to see if similar results can be obtained simply by chance (see results from randomization from publication V).

As a solution, an automatic method for finding cluster enriching functional classes was developed and tested with three different hierarchical clustering methods for the analysis of the Hughes data set (publication V). The results from average method were analyzed more in detail. Bonferroni corrected p-value limit 0.05 (Dudoit, et al., 2000) was used as a condition for stopping the search. The method identified 116 clusters with mainly analyzed MIPS classification and 176 clusters with SGD classification. The results are given in their estimated order of goodness, which is obtained from the log-p-values of the functional class with the best score.

The first 50 clusters from the obtained results were analyzed more in detail (table 1 in publication V and 23 best clusters shown here in table 2). These clusters show very significant correlations as can be seen from the obtained p-values. As an example, cluster 9 includes all eight histone proteins and only the histone proteins with bonferroni corrected log-p-value ~22 (table 2). Strong co-regulation of the histone proteins has been reported earlier (Eisen, et al., 1998) but not in the analysis of this data set (Hughes, et al., 2000). Other similar findings are cluster number 2, including 28 out 31 respiratory proteins and cluster 12, including almost exclusively ribosomal complex proteins. More examples can be seen in the table 1 in publication V.

### Table 2 -23 best clusters from average method with MIPS classes from the cluster search method

<table>
<thead>
<tr>
<th>ordinal number</th>
<th>most enriched MIPS gene class</th>
<th>log-p-value</th>
<th>bonf. corr. log(p)</th>
<th>observe d</th>
<th>clust. size</th>
<th>class size</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>cytoplasmic ribosomes</td>
<td>102.5676</td>
<td>99.7780</td>
<td>78</td>
<td>159</td>
<td>108</td>
</tr>
<tr>
<td>2</td>
<td>respiration chain complexes</td>
<td>61.0379</td>
<td>58.2483</td>
<td>28</td>
<td>41</td>
<td>31</td>
</tr>
<tr>
<td>3</td>
<td>mitochondrion</td>
<td>57.3513</td>
<td>54.5617</td>
<td>75</td>
<td>142</td>
<td>305</td>
</tr>
<tr>
<td>4</td>
<td>amino acid metabolism</td>
<td>55.4905</td>
<td>52.7099</td>
<td>75</td>
<td>259</td>
<td>171</td>
</tr>
<tr>
<td>5</td>
<td>amino acid biosynthesis</td>
<td>55.4849</td>
<td>52.6953</td>
<td>59</td>
<td>244</td>
<td>98</td>
</tr>
<tr>
<td>6</td>
<td>mitochondrial ribosomes</td>
<td>46.2851</td>
<td>43.4955</td>
<td>32</td>
<td>101</td>
<td>46</td>
</tr>
<tr>
<td>7</td>
<td>rRNA transcription</td>
<td>33.5036</td>
<td>30.7140</td>
<td>46</td>
<td>274</td>
<td>99</td>
</tr>
<tr>
<td>8</td>
<td>cytoplasmic ribosomes</td>
<td>25.8871</td>
<td>23.0975</td>
<td>15</td>
<td>15</td>
<td>108</td>
</tr>
<tr>
<td>9</td>
<td>nucleosomal protein complex</td>
<td>25.2346</td>
<td>22.4450</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>10</td>
<td>26S proteasome</td>
<td>23.1373</td>
<td>20.3477</td>
<td>12</td>
<td>21</td>
<td>28</td>
</tr>
<tr>
<td>11</td>
<td>Glycolysis and gluconeogenesis</td>
<td>21.1692</td>
<td>18.3796</td>
<td>10</td>
<td>13</td>
<td>28</td>
</tr>
<tr>
<td>12</td>
<td>cytoplasmic ribosome large subunit</td>
<td>20.5629</td>
<td>17.7734</td>
<td>11</td>
<td>12</td>
<td>63</td>
</tr>
<tr>
<td>13</td>
<td>purine ribonucleotide metabolism</td>
<td>20.4238</td>
<td>17.6342</td>
<td>9</td>
<td>9</td>
<td>33</td>
</tr>
<tr>
<td>14</td>
<td>mitochondrion</td>
<td>19.3933</td>
<td>16.6038</td>
<td>19</td>
<td>24</td>
<td>305</td>
</tr>
</tbody>
</table>
The table is a part of the results obtained in publication V. The table shows the ordinal number of the each obtained cluster shown in the figure 2 in publication V, functional class that was most enriched in the cluster, and the log10-p-value for the most enriched class with and without Bonferroni-correction. The table shows also number of class members in the cluster (observed), size of the cluster (clust. size), and size of the functional class. Notice that clusters that include only or almost only functional class members (for example 8, 9, 12, 13, 15, 19) and the clusters that include the whole or almost the whole functional class (2, 6, 9). An especially good example is cluster 9 as it includes all and only the histone complex members. It is worthwhile to notice the extremely small obtained p-values (for example, log-p-value 12 would here refer to p-value $10^{-12}$). The whole table is shown in the original publication V and in its supplementary information.

Detailed analysis showed redundancy among the obtained functional groups, as many of them were associated to the same cellular processes. This was presented by rough manual grouping of the clusters according to enriched functional classes (table 2, publication V). Grouping gives a rough overview of the cellular processes that show response to created mutation stresses. Results were also in good agreement with earlier manual analysis of enriched functional groups although earlier manual work analyzed several different cluster trees (table 3 in publication V). Groups were further enlarged by forming a group of energy production associated clusters, group of protein synthesis associated clusters, and a group of clusters associated with cell cycle, differentiation, or nucleotide processes. These where in turn used in the visualization to highlight the placement of these processes in the cluster tree (figures 2 and 3 and table 2 in publication V). The obtained clusters enable an analysis of distribution of different functions on top of the cluster tree similarly to previous work with SOM (publication III). The detailed analysis of these distributions can be found in the article.

As was previously stated, an automated method can be easily tested with randomized data sets to see whether similar results could be obtained. Testing was done by randomizing the connections between the genes in the table including functional classifications and the genes in the table of expression profiles. Results included the log-p-values for all the clusters in the cluster tree separately in each randomization. Results from randomization were totally different (average 99th of raw log-p-values ~4) from the obtained results (mainly analyzed raw log-p-values ranged from 7 – 102).

It has been previously pointed out that the correlation of clustering results with gene classes could be used to evaluate different clustering and preprocessing methods (Gibbons and Roth, 2002, Oja, et al., 2002). Such evaluation is natural as the genes with similar function or location are expected to be similarly regulated. Here the obtained clusters were used to compare three clustering methods (average and complete method and Ward’s minimum variance method). Clusters that match each other in different
clustering results were compared to see which one obtained stronger enrichment. This was analyzed by creating two-dimensional scatter plots to see if a stronger enrichment can be seen regularly in one of the clustering methods (see figure 7 in publication V). It was observed that there was a trend between average and complete in favor of better results with average clustering. A similar weaker trend was observed between Ward’s method and complete. The results from comparison of Ward’s method and average, on the other hand, did not show considerable differences. The obtained results suggested the joint analysis of results from Ward’s method and average.

5.4. Detailed analysis of the cluster content

All the work described before has extensively used gene classes to select cluster areas. Gene classes have been also used for more detailed analysis. The method simply calculates the developed correlation score for correlation of cluster with all analyzed gene classes and reports the obtained log-p-value and bonferroni corrected log-p-value (Dudoit, et al., 2000) as a sorted list starting with the strongest enrichment (see table 3). Instead of limiting reporting to this, other information is given like the size of the gene class in the cluster, size of the class in whole data set and expected number of occurrences for the gene class in question from random sampling (Törönen unpublished).

The main idea has been directing the analysis to the most significantly grouped functional classes and presenting them in a sorted order so that the important features can be easily seen from the list. Remember that usually the number of different gene classes viewed in this way can be several thousands making the analysis without the sorting impractical. Table 3 shows an example output from the discussed analysis (Törönen unpublished). The cluster is obtained with Discriminative Clustering (Sinkkonen and Kaski, 2000) and it can be clearly seen that the cluster is associated with energy production associated metabolism.

The importance of the obtained log-p-value statistics should be evaluated. The additive information included in the table enables one to observe whether the most of the genes in the cluster belong to a quite common functional class. This would be the case where the specificity is high although the sensitivity is not equally high. As an example of these cases table 3 shows functional classes of metabolism and energy. Approximately half of the genes in the cluster belong to these classes (~60% belong to metabolism, ~50% belong to energy). These work as a description of the whole cluster giving an overview of the cluster content. Still it would be naïve to estimate the behavior of the whole metabolism class on basis of the obtained result.

The other case can be where most of the class members of a small class belong to a bigger cluster. This, on the other hand, would be a case where the sensitivity is high but the specificity is not equally high. As an example of such cases, table 3 shows classes ‘tricarboxylic-acid pathway’ and ‘glycolysis and gluconeogenesis’. This forms a good starting point for more detailed analysis of these functional classes, as most of the functional class in question is included to the cluster in question. The drawback is of
course the fact that these correlations cannot be used to define the content of the whole cluster like the metabolism and energy classes. Despite the possible weaknesses in the correlation (low specificity or sensitivity) all the previous cases are found to be statistically highly significant and can be helpful during the data analysis. It should be noted how the cases of low specificity and high sensitivity and cases of high specificity and low sensitivity complement each other giving different information about the obtained cluster.

### Table 3 –An example output from the cluster analysis method. 20 Best scoring MIPS functional classes for clusters created with the Eisen data set.

<table>
<thead>
<tr>
<th>functional class</th>
<th>-\log_{10}(P)</th>
<th>P-value</th>
<th>size of the class</th>
<th>observed number of genes</th>
<th>expected number of genes</th>
<th>STD of expected number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Energy</td>
<td>82.54808</td>
<td>0</td>
<td>169</td>
<td>107</td>
<td>14.28943</td>
<td>3.491061</td>
</tr>
<tr>
<td>C-compound and carbohydrate utilization</td>
<td>41.25301</td>
<td>0</td>
<td>158</td>
<td>74</td>
<td>13.35935</td>
<td>3.383629</td>
</tr>
<tr>
<td>C-compound and carbohydrate metabolism</td>
<td>27.68014</td>
<td>0</td>
<td>263</td>
<td>79</td>
<td>22.2374</td>
<td>4.264758</td>
</tr>
<tr>
<td>respiration</td>
<td>25.11711</td>
<td>0</td>
<td>68</td>
<td>39</td>
<td>5.749593</td>
<td>2.26275</td>
</tr>
<tr>
<td>glycolysis and gluconeogenesis</td>
<td>25.07987</td>
<td>0</td>
<td>29</td>
<td>26</td>
<td>2.452033</td>
<td>1.48968</td>
</tr>
<tr>
<td>Metabolism</td>
<td>23.71709</td>
<td>0</td>
<td>705</td>
<td>127</td>
<td>59.60976</td>
<td>5.240719</td>
</tr>
<tr>
<td>mitochondrial organization</td>
<td>18.75441</td>
<td>0</td>
<td>303</td>
<td>73</td>
<td>25.61951</td>
<td>4.535737</td>
</tr>
<tr>
<td>tricarboxylic-acid pathway</td>
<td>16.20465</td>
<td>0</td>
<td>17</td>
<td>16</td>
<td>1.437398</td>
<td>1.143372</td>
</tr>
<tr>
<td>glyoxylate cycle</td>
<td>5.383624</td>
<td>0.000004</td>
<td>5</td>
<td>5</td>
<td>0.422764</td>
<td>0.621602</td>
</tr>
<tr>
<td>fermentation</td>
<td>5.355452</td>
<td>0.000004</td>
<td>14</td>
<td>8</td>
<td>1.18374</td>
<td>1.03823</td>
</tr>
<tr>
<td>pentose-phosphate pathway</td>
<td>5.082534</td>
<td>0.000008</td>
<td>8</td>
<td>6</td>
<td>0.676423</td>
<td>0.78579</td>
</tr>
<tr>
<td>organization of cytoplasm</td>
<td>4.761521</td>
<td>0.000017</td>
<td>471</td>
<td>64</td>
<td>39.82439</td>
<td>5.430369</td>
</tr>
<tr>
<td>other energy generation activities</td>
<td>4.637107</td>
<td>0.000023</td>
<td>9</td>
<td>6</td>
<td>0.760976</td>
<td>0.833287</td>
</tr>
<tr>
<td>ionic homeostasis</td>
<td>4.167948</td>
<td>0.000068</td>
<td>94</td>
<td>20</td>
<td>7.947967</td>
<td>2.645896</td>
</tr>
<tr>
<td>Transport facilitation</td>
<td>3.575733</td>
<td>0.000266</td>
<td>186</td>
<td>30</td>
<td>15.72683</td>
<td>3.648826</td>
</tr>
<tr>
<td>Cellular organization</td>
<td>3.433077</td>
<td>0.000369</td>
<td>1806</td>
<td>173</td>
<td>152.7024</td>
<td>6.097464</td>
</tr>
<tr>
<td>mitochondrial transport</td>
<td>3.421321</td>
<td>0.000379</td>
<td>54</td>
<td>13</td>
<td>4.565854</td>
<td>2.022303</td>
</tr>
<tr>
<td>transport ATPases</td>
<td>3.377159</td>
<td>0.000422</td>
<td>35</td>
<td>10</td>
<td>2.95935</td>
<td>1.634524</td>
</tr>
<tr>
<td>Amino-acid metabolism</td>
<td>3.289599</td>
<td>0.000513</td>
<td>158</td>
<td>26</td>
<td>13.35935</td>
<td>3.383629</td>
</tr>
</tbody>
</table>

The table is an example of the output from the method that was used to evaluate the obtained clusters. Functional classes associated to analysed clusters are being presented in the order of their log-p-values. Additional information is shown in order to demonstrate what type of correlation is in question. The cluster included 208 genes and it was obtained using Discriminative Clustering (Sinkkonen and Kaski, 2000; Toronen et al. unpublished work). Notice the reporting of the groups that constitute a large part of the cluster (ENERGY, METABOLISM) and the groups that are almost totally included by the cluster (tricarboxylic-acid pathway, glycolysis and gluconeogenesis). Note that this table differs from table 2 as here the most significantly correlating functional classes are shown for one cluster whereas in the table 2 several clusters are shown with the highest scoring functional class for each of them.

### 6. Discussion

Although gene expression has great potential for enhancing the understanding of biological systems and gene functions, there are also weaknesses in the obtained data. The reliability of a single gene expression measurement is often questionable. Also the size of the data creates problems for the direct analysis and increases the possibility of
obtaining false positive results. Here the author has presented ways to improve the analysis of gene expression data. The main emphasis is on two issues: use of clustering and visualization methods to divide data into analyzable subgroups, and evaluation of clustering and visualization results using the available classifications of the genes to find the important features and to enhance the analysis.

The first step in the analysis of gene expression data is the reliable estimation of regulation of the gene. This forms the basis on top of which the rest of the analysis is build upon. In our first in house analysis, we noticed that the normal ratio measurement is too unreliable. We developed a very simple and crude correction method for this problem that still was able to fix the problem (publication I, IV). Other groups have also found the presented method useful (Hiltunen, et al., 2002). The drawback of the idea is the selection of the suitable constant. Here, more advanced methods using statistical modeling of the data (Durbin, et al., 2002, Newton, et al., 2001) could become helpful. It should be also noted that variance-stabilizing functions (Durbin and Rocke, 2004, Huber, et al., 2002), local estimates of the data variance (Kepler, et al., 2002), and other data modeling methods aim also for this same purpose. The large list of available methods actually raises the question of detailed comparison of the weaknesses and strong points for each method. Although it is easy to get fascinated with advanced clustering, classification, and other data mining methods, these will not help much if the basis for the data production is weak.

Publication I presented screening of genes that show regulation in the MK-801 treatment. A more detailed evaluation of the analysis points to two weaknesses. First the normalization was done using only one housekeeping gene. As the reliability of the measurements is questionable (see chapter 2.2) it is very likely that the observed measurements for one housekeeping gene will have some amount of error. Therefore the use of several housekeeping genes would have given a more reliable estimate. Also the housekeeping genes could have worked as controls against each other as the reliability of the constant expression of one housekeeping gene could have been evaluated using the others to test it. The other critical question is the magnitude of the regulation that was very weak for all the analyzed genes. A reliable estimation of low-level regulation would require a very large number of replicates. Here the housekeeping genes could have worked as negative controls. Still it should be noted that the confirmed regulations were also at the magnitude smaller than 2 pointing that the important changes in the gene expression in the neuronal tissues are often at very small magnitude.

Our first work in the analysis of expression profile level showed that the clustering and the visualization with SOM and Sammon’s mapping enable us to see the important regulatory trends in the data set (publication II). It was also shown with the example cluster areas that the genes grouped inside single map units show very conserved biological functions. The only exception was the area in the SOM map where map units did not show any significant regulation. We actually visualized, in a later further analysis of the data set, the distribution of the selected genes associated with proteasome complex in order to visualize the distribution of assumed genes with no regulation in the test situation (data not shown). The obtained visualization showed that these genes fell
randomly into the member map units of the middle cluster in the SOM map (clustered SOM map units in the middle of the Sammon’s mapping in figure 2c in publication II). This method could be later used to manually filter low-variance genes. Here, proteasome genes actually function as an experiment specific housekeeping gene set. The same idea could be applied to other data sets in order to estimate the variance among the genes that are expected to have no significant signal. A similar idea of negative control formed by the measurements that should not have significant variance was later used here in the analysis of the Hughes data set to estimate the cutoff for low-variance genes (publication V).

Our second work with SOM took the idea of using the gene classes further and it used them extensively to associate the map regions to different cellular processes (publication III). We also showed that the SOM algorithm could be used with gene expression data as a preliminary weak clustering step from which a manual clustering of the interesting cluster areas can be produced. Another group pioneered at the same time the use of SOM in the analysis of gene expression data (Tamayo, et al., 1999). The emphasis in their analysis is more in the preliminary clustering created by SOM and the further processing like clustering neighboring SOM nodes was not done. Such usage makes SOM more a clustering method and does not take the full advantage of the network structure. Also the visualization of the distances between the SOM nodes is crucial as they can vary a lot. As an example of the varying distances between nodes one can see the visualization in publication II in figure 2C. This information would be lost without the Sammon’s mapping or U-matrix type visualizations. Another competing method for clustering and visualization is hierarchical clustering (Eisen, et al., 1998). We showed in publication III that the reliability of the visualization produced by hierarchical clustering is weaker than what can be obtained with SOM. In the same work we show easy and intuitive visualization of the distribution of the functional classes over the SOM map. Similar presentation of the class distributions would be harder to show on top of the cluster tree from hierarchical clustering. Still, one possible way for doing the analysis of functional class distributions is shown in publication V.

Earlier work has shown that the functional classification of genes could be used to find interesting cluster areas from SOM. Still there are often many possible choices for defining the cluster borders. If many overlapping clusters are defined the analyst is left to decide which ones to select for detailed analysis and also with manual analysis of content of several clusters. To overcome these two problems author defined a method for analysing and evaluating cluster regions (publication IV). The method simply reported the analysed functional classes in the estimated order of importance showing also additive information for more detailed evaluation of the obtained result (see table 3 for example). This way, the analyst can base selection of clusters to reported statistical significances of the observed enrichments. Notice that usually the reports showed several tens of classes for each cluster so that an accidental enrichment of a single functional class should not affect the selection much. Still, it should be noted that obtained p-values are not useful in the same way as normal p-values as the analysed clusters have been selected from a larger pool of many clusters using the p-value as a selection method. Selection naturally increases the probability of obtaining a more significant p-value. The
resulting bias could be corrected using bonferroni correction type methods. The presented method has played an important part in the later research in our group (Knuuttila, et al., 2004, Lähteinen, et al., 2004).

A step further in the search of interesting cluster areas was to automate the cluster search method. A method was developed here that used another popular clustering method, hierarchical clustering, to create a starting clustering as a basis to search clusters with strong correlation between clusters and functional classes (publication V). The created search method was able to find very significant class-cluster correlations. The reported results would have been hard to find by using other cluster tree analysis methods like cutting the tree at the selected level or by analysing the sorted list of genes using short descriptions. It was also shown that the obtained results could not come from random sampling of class data set. Besides finding interesting clusters the used method formed the basis for the comparison of different hierarchical clustering methods. The obtained results proposed a joint analysis of clustering results from average linkage and Ward’s minimum variance method. Still, the analysis was limited to analysis of average method to make the work more compact. Combining different clustering results for analysis can cause problems and one solution for combining the results was presented in publication V. The method first identifies clusters with best correlations and then selects the cluster with better enrichment for analysis.

It should be noted that our collaborating group has developed methods that aim at the same goal of finding the localized cluster areas showing significant enrichment for some functional gene group. Their method, discriminative clustering (Sinkkonen and Kaski, 2000), looks for data partitioning similar to k-means that maximizes the correlation of the local areas with the gene functional classes. These two methods should not be considered as competing but rather complementing methods as they could be used in parallel or in series to direct each other to more reasonable results. The benefit of the presented method to discriminative clustering is that the important functional classes do not have to be selected beforehand. Instead they are reported in the results with the obtained cluster. Therefore the presented method could be used to select the functional classes to be used with discriminative clustering. Similarly the discriminative clustering could be used to lessen the impact of dimensions that include only error to the hierarchical clustering process.

Unpublished work presented one way of simplifying the analysis of biological themes among the genes in the cluster. The presented method sorted the enriched classes according to the statistical significance of the enrichment. Similar principles have been also presented in the literature (Hosack, et al., 2003) with EASE software. The difference between EASE and the current method is the presented measure of the correlation. The one used here can identify separately both significant over-representation and under-representation of the functional class. Although the under-representation was not much used in the current work, it can still be important in some applications. For example in the analysis of the enrichment of the motifs in the up-stream area of up-regulated genes a lack of certain motif can signal a binding site of an expression repressor.
Throughout our gene expression analysis, the functional annotation and classification of genes to classes based on their biological function, membership of a protein complex, placement inside the cell, membership of a pathway, chemical function, or existence of some keywords in their functional description has played a very important role. It should be noted that categorization of genes is bound to do some injustice to the actual associations of the genes to the categorized functions as it is obvious that the genes are associated with different degrees to these functions. Genes from the same functional class can also be involved in either up or down-regulation of the function making them behave differently. Many of the genes also usually lack the functional description or it can be erroneous. Still, even with their weaknesses, multiple levels of parallel gene classifications usually direct the analysis of the gene group to the correct direction, as most of the classifications are usually correct. The use of functional annotations to evaluate the results gives some benefits:

1) Analysis is directed to areas in the data space where there is a significant group. Even though most of the analysis usually looks for groups of similarly behaving genes that can be grouped to a cluster, it is usually of greater interest to analyze regions of data space where a certain biological function is associated with the presented expression profile.

2) Analysis of obtained clusters/gene groups can be made more intuitive. Usually the analysis of gene groups is based on biological knowledge of the obtained group but this approach, although it always should be the last step to evaluate the obtained result, has drawbacks. First, it is hard to be aware of functions for several genes and to keep up with the latest findings for the genes. Equally, it might not be immediately obvious what are the connecting features that explain the co-regulation of the obtained genes. Last, it is hard to estimate which of the observed common functions are actually statistically significantly grouped to the cluster. As a naïve example, one can observe a group of genes where most of the genes belong to same functional class but which still is not significant if most of the genes in the rest of the data set belong to the same group.

3) Several genes from the same functional group give a more reliable estimate about the regulation of the biological function coded by these genes. As the reliability of an individual observation is very questionable usually in the gene expression data, the observation that many genes from the same functional group present the same regulation, makes the hypothesis, that the function in question is actually regulated in the presented manner, more reliable.

4) Different groups can easily be compared for similarities/dissimilarities in the behavior in the expression data. Analysis of the shown functional groups can be taken a step further by looking at what biological processes show similar reliable regulation. This helps to create a more comprehensive image about biological situation in the test situation.

A drawback in the use of functional annotation in the evaluation process is possible errors and missing functional classifications in the used functional classification. Also, the biological processes that are regulated by one rate-limiting enzyme that is also the only
gene that is regulated can go unnoticed in the analysis. Still, it should be noted that the latter might easily go unnoticed also in the manual analysis unless there is specific information that the gene in question is a rate-limiting enzyme. Also when the information on a rate-limiting enzyme is available it could also be included to the analysis just like the functional annotations.

Besides using gene classifications, we also used keyword searches to analyze short functional descriptions. Use of keywords has also been modified in our later work by using associated Swissprot (Apweiler, et al., 2004, Boeckmann, et al., 2003) entries for each analyzed genes (Knuuttila, et al., 2004, Lähteinen, et al., 2004). Swissprot is a database presenting detailed information about the well-annotated protein sequences. Each Swissprot entry included keywords that were taken then into the analysis and used as functional categories. Also, other data sets could be used similarly to confirm that the obtained cluster is not a result of random clustering. Protein-protein interaction data can be used similarly to confirm the obtained clustering. The association of the gene expression profiles and occurrence of potential transcription binding sites have been compared similarly as was done here with gene expression data and functional class categories (Brazma, et al., 1998, Segal, et al., 2003).

The major problem with most biological data sets is the selection of true signal from the pool of noise signals often found in large data sets. Combination of different data sets will therefore probably play an even more crucial role in the future of the biological data analysis as it can enable the combination of the true signal from several sources. Already, the combination of different cancer data sets have revealed common features throughout different cancer types (Rhodes, et al., 2004) and similarly the comparison of the expression data from different species has revealed the conserved behavior of certain functional gene groups (Rhodes, et al., 2004). Also, the parallel use of several methods for annotation of unknown genes has been shown to be useful (Marcotte, et al., 1999).

All the previous methods are from the area of data mining and they could be applied equally to any other large data set. Similar analysis situations come in biochemistry from protein-protein interaction data, analysis of protein sequences, and/or structures and finding important factors for gene function. An interesting area with complex problems is also the analysis of biological text data sets for extracting information like the function of the associated genes or the interactions between genes. Also, areas like diagnostics where some potential applications of gene expression data analysis lie could apply the used analysis methods.

7. Conclusions

The presented work aimed at developing methods for our research group to enhance gene expression data analysis. We were able to lessen the effect of additive error in our analysis steps with simple modification to the ratio calculation. We tested the usefulness of the SOM algorithms to the clustering and visualization of the gene expression data. We showed that the results were similar to previous analysis and that especially the
analysis of the distribution of the functional classes in the SOM can help to analyse the data. Furthermore, some methods, especially concentrating on the analysis of the enrichment of functional classes in the clusters, and on the search of clusters showing strong correlations with analysed functional classes, were developed. The presented novel methods should enhance the analysis of large data sets as they can emphasize the details that are informative for the analysis process.
8. References


