HUMBOLDT-UNIVERSITÄT ZU BERLIN

MATHEMATISCH-NATURWISSENSCHAFTLICHE FAKULTÄT I
INSTITUT FÜR BIOLOGIE

DIPLOMARBEIT

ZUM ERWERB DES AKADEMISCHEN GRADES
DIPLOM-BIOLOGIN

Effects of Oncogenic Ras on Gene Expression: Clustering of Microarray Data and Screening for Potential Serum Response Factor Targets

vorgelegt von
Karoline Faust
geb. am 28. März 1980 in Berlin

Gutachter:
Prof. Dr. H. Herzel
PD Dr. C. Sers

Berlin, 15. September 2005

Eigenständigkeitserklärung

Hiermit versichere ich, dass ich die vorliegende Diplomarbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Karoline Faust Berlin, den 15. September 2005

Zusammenfassung

Humane Ras-Onkogene spielen eine wichtige Rolle bei der Entstehung vieler Krebsarten. Die Aufklärung der durch aktivierte Ras-Proteine gesteuerten Signalwege und deren Einfluss auf die Genexpression ist daher von besonderer Bedeutung.

In der vorliegenden Arbeit wurde ein Microarray-Experiment ausgewertet, mit dem die Wirkung eines induzierbaren mutierten Ras-Proteins auf die Genexpression in embryonalen Rattenfibroblasten untersucht wurde. Das Expressionsniveau von 311 Genen wurde über eine Zeitspanne von 8 Tagen zu insgesamt 15 verschiedenen Zeitpunkten gemessen. Die Messungen erfolgten in den ersten 5 Tagen nach Induktion des Ras-Proteins, anschliessend für weitere drei Tage nachdem die Induktion beendet worden war.

Im ersten Teil der Arbeit wurden die Microarraydaten mit verschiedenen Northern blot und Real-time PCR Experimenten, die zur Verifizierung der Arraydaten anhand von 10 Genen durchgeführt worden waren, verglichen. Für jedes Experiment wurde im Vergleich zum Microarray Experiment der Spearman Koeffizient errechnet.
Im zweiten Teil wurden Genexpressions-Vektoren von 82 signifikant differentiell exprimierten Genen unter Verwendung von hierarchischen (agglomeratives Nesting) sowie partitionierenden (Kmeans, SOM, PAM) Verfahren geclustert. Das Problem der Wahl einer angemessenen Clusterzahl wurde mit Hilfe von Clusterindizes angegangen. Unter Verwendung der verschiedenen Methoden konnten vier Gengruppen identifiziert werden, die eine deutliche Änderung ihrer Expressionswerte nach Ras-Induktion und erneut nach Beendigung der Induktion zeigten.

Eines der vier Cluster mit insgesamt 25 Genen, deren Expression in Abhängigkeit von Ras stimuliert wurde, zeigte ein Expressionsprofil, welches über den Beobachtungszeitraum von 8 Tagen dem Profil des Serum Response Factors (SRF) glich. Ausserdem enthielt dieses Cluster drei bekannte Zielgene des Transkriptionsregulators SRF. Daher wurde im dritten Teil der Arbeit untersucht, ob in diesem Cluster weitere potentielle SRF-Zielgene liegen. Hierfür wurden Up- und Downstream-Bereiche aller Gene dieses Clusters extrahiert. Im nächsten Schritt wurden konservierte Bereiche durch Alignments von Ratten-, Maus- und menschlichen Sequenzen aufgespürt. Innerhalb dieser Bereiche wurde dann eine Suche nach möglichen SRF-Bindestellen mit Matrizen der Datenbank T-Reg durchgeführt. Mit dieser Vorgehensweise konnten drei bekannte SRF-Zielgene (cpg21, PDGF-A, junB) bestätigt und ein neuer Kandidat (MKP-3) vorhergesagt werden.
MKP-3 ist Teil einer negativen Rückkoppelungsschleife, die aktiviertes MAPK ausschaltet. Seine Regulation ist von besonderem Interesse, da es ein potentieller Tumorsuppressor ist. Die vorliegende Arbeit deutet darauf hin, dass aktiviertes Ras MKP-3 über SRF hochreguliert.

Summary

Human Ras oncogenes play an important role during the formation of many cancers. Therefore, elucidation of signaling pathways downstream of Ras and their influence on gene expression is of special interest.

The present work analyzes a microarray experiment performed for studying the effects of inducible oncogenic Ras on gene expression in rat embryonal fibroblasts. The expression levels of 311 genes were investigated during 8 days with a total of 15 time points. Measurements were conducted during 5 days after induction of the Ras protein and for additional three days after abrogating Ras induction.

In the first part of the thesis, microarray-derived expression values were compared to verification data obtained via Northern blot or real-time PCR analysis for 10 selected genes. For each experiment a Spearman correlation coefficient was calculated in comparison with the microarray data.

In the second part, gene expression vectors of 82 significantly differentially expressed genes were clustered using hierarchical (agglomerative nesting) as well as partitional methods (kmeans, PAM, SOM). The problem of choosing the right cluster number was tackled with the help of cluster indices. Using the various methods, 4 clusters were identified which show a significant alteration in their expression level after Ras induction and again after cancelling the induction.

One of the four clusters (B) harbouring 25 genes, whose expression is stimulated in response to Ras, exhibited an expression profile similar to the profile of the transcription factor serum response factor (SRF) within the 8 days of measurement. In addition, this cluster was found to harbour three known target genes of SRF. Therefore, this cluster was analyzed for further potential SRF targets during the third part of the thesis. For this purpose, regions up- and downstream of the genes belonging to this cluster were retrieved. As a next step, conserved regions where detected by an alignment of rat, mouse and human sequences. Within these regions, a search for potential SRF binding sites was performed using matrices obtained from the database T-Reg. By this procedure, the three known SRF targets could be confirmed (cpg21, PDGF-A, junB) and one new target (MKP-3) was predicted.

MKP-3 is part of a negative feedback loop that switches off activated MAPK. Its regulation is of particular interest, since MKP-3 is a potential tumor suppressor. This study suggests that MKP-3 is up-regulated upon Ras activation via SRF.

Danksagung

Hiermit möchte ich mich insbesondere bei Frau Dr. Sers bedanken, ohne deren ständige Unterstützung die vorliegende Arbeit nicht zustande gekommen wäre.
Herrn Prof. Herzel danke ich für die freundliche Betreuung und die geduldige Hilfestellung bei schwierigen Konzepten.
Vielen Dank an Steffen Grossmann für die Betreuung während der SRF Targetsuche, für den Zugang zu T-Reg und den lokalen Ensembl-Versionen, die Ets- und SRF-Matrixlisten sowie die oft erprobte Bereitschaft, bei Problemen weiterzuhelfen.
Nils Blüthgen, dem zuverlässigen Retter in der Not, ein herzliches Dankeschön für die vielen Diskussionen und guten Tipps.
Jana Keil möchte ich danken für die Betreuung in den Anfangsmonaten und ihr freundliches Entgegenkommen bei allen Fragen und Problemen.
Oleg Tchernitsa danke ich für die Einführung in gängige Software zur Microarray-Analyse und die Bereitstellung der GenBank-IDs.
Irina Nazarenko, Minh-Cam Ha-Thi, Rula Abdul-Ghani und den anderen Mitgliedern der Arbeitsgruppe von Dr. Sers und Prof. Schäfer ein herzliches Dankeschön für die angenehme Arbeitsatmosphäre und freundlich gewährte Hilfe.

Abbreviations

AC agglomerative coefficient

bp base pair (1kb = 1000 bp)

CNB conserved non-coding sequence block

DUSP dual specificity phosphatase (alias MKP)

DBTSS DataBase of Transcriptional Start Sites

ERK extra-cellular signal regulated kinase (alias MAPK)

EPD Eukaryotic Promoter Database

FP false positive

FN false negative

$\mathrm{IC}_{\mathrm{rel}}$ relative information content

IEG immediate early gene

IPTG isopropyl-1-thio-β-D-galactosidase

IR-4 inducible Ras clone 4

Lox Lysyl oxidase

MAPK mitogen-activated protein kinase

MKP MAPK phosphatase

GO gene ontology

PAM partitioning around medoids

PCA principal component analysis

PDGF-A A chain of the platelet derived growth factor

P-ERK phoshporylated (activated) ERK (alias P-MAPK)

PFM position specific frequency matrix

PI3K phosphoinositide-dependent protein kinase

PSCM position specific count matrix

PSSM position specific score matrix

RefSeq NCBI Reference Sequence

PCR polymerase chain reaction

SOM self-organizing map

SRE serum response element

SRF serum response factor

TCF ternary complex factor

TFBS transcription factor binding site

TSS transcription start site

Contents

1 Introduction

1.1 Biological Background of Ras Signaling

Why are Ras proteins crucial for the development of many cancers? In normal cells, Ras is transferred from an inactive GDP-bound state to an activated GTP-bound state in response to growth factor signals. The growth signal is switched off again by GTPase activating factors, which enhance the intrinsic capability of Ras to hydrolyze GTP. In human tumors, mutations within the Ras genes result in the generation of oncogenic Ras proteins, which prevent hydrolysis of GTP by their altered conformation, leading to a continuous transmission of growth stimulatory signals [CAMPBELL ET AL. 98].
Thus, Ras acts as a molecular switch, which sets the course for proliferation. It does so by its influence onto various intracellular processes such as protein modification, transcription and translation [CAMPBELL ET AL. 98]. Signal transduction emerging from Ras proteins does not follow a linear path, but forms a complex network. In the following, a small, but well characterised part of this network, called the MAPK-cascade, is described in detail.

If a growth factor binds to its receptor, the transition of Ras to its active, GTP-bound state is mediated via adaptor proteins. Then, the signal is passed on through three layers of kinases, the first being the serine/threonine kinase Raf. Activated MEK, the kinase on the second level, phosphorylates the third level kinases p42/p44 MAPK (in the following named MAPK), which in their active state enter the nucleus. The extend of MAPK accumulation within the nucleus depends on the nature of the stimulus. As Volmat et al. demonstrated, only a mitogenic stimulus allows long term activation of MAPK and consequently the transcription of nuclear anchors, which fix MAPK in the nucleus [VOLMAT ET AL. 01].

Next, the activation of the transcription factor Elk-1 by activated MAPK initiates the transcription of target genes. Not only Elk-1 but also a variety of other transcription factors is activated by Ras via additional pathways, notably SRF, ATF2, Jun and NF-κB [CAMPBELL ET AL. 98].

An important group of Ras target genes are the dual specificity phosphatases (DUSPs). They play a key role in the anchoring and inactivation of MAPK in the nucleus (DUSP1, DUSP2 [VOLMAT ET AL. 01]) as well as in the dephosphorylation of MAPK in the cytoplasm (DUSP6, DUSP9 [CAMPS ET AL. 98]) and therefore can work as tumor suppressors [FURUKAWA ET AL. 03].

In the literature, DUSPs often are referred to as MAPK phosphatases (MKPs) because of their specificity for mitogen-activated kinases. DUSP6 for example is also named MKP-3, which is the name used throughout this work. DUSP1 and MKP-1 are synonyms as well as DUSP9 and MKP-4, DUSP5 is also known as cpg21. An overview of different names for a number of DUSPs is given in [CAMPS ET AL.00].

1.2 Microarray Experiment

During recent years, the influence of Ras oncogenes on gene expression was analyzed quantitatively by several groups. The first description of a genome-wide expression profile of Ras-transformed rat fibroblasts as compared to immortalized rat fibroblasts was published in 2000 [ZUBER ET AL.00]. For nearly half of the genes shown to be differential in this cell culture model, differential expression was verified with Northern blots. In addition, the same group established a cell line derived from the immortalized rat fibroblasts containing an inducible HRAS (G12V) (IR-4 cell line). After induction of oncogenic HRAS using IPTG, not only the morphology of IR-4 cells changes from flat to spindle-like but the cells also acquire the capacity to grow anchorage-independently, a typical characteristic of transformed cells [SERS ET AL.02].

In order to study the time course of gene expression in response to oncogenic Ras, 311 genes derived from the initial genome wide expression profiling were analyzed using a series of customized Ras target specific gene arrays (Tchernitsa&Sers, unpublished). mRNA from non-induced and induced IR-4 cells was collected after 0, 10 min, 30 min, 60 min, 2 h, 6 h, 12 h, 24 h, 48 h, 72 h, 93 h and 120 h. After 5 days (120 h), the inducer IPTG was removed from the cell culture medium and mRNA was collected at three further time points: 144 h, 168 h and 192 h. All RNA samples were subsequently used for the generation of cDNA and hybridized to the array.

For the generation of the customized Ras-target gene array, a specific oligomer of 70 base pair length was designed for each gene and checked for unwanted homologies using BLAST. Because mRNA is transcribed into cDNA before usage, the oligomer is complementary to parts of the probe and can hybridize with it. The process of reverse transcription of the probe mRNA also serves for labeling, either with a red (Cy5) or green (Cy3) fluorescent. For each time point, a dye swap design was employed such that on the first of two arrays the probe taken from non-induced cells was colored red and the probe from induced cells green whereas on the second array the coloring of the probes was exchanged. The target oligomers were spotted five times for each gene on poly-l-lysine treated glass slides, which also contained 20 different house keeping genes and positive and negative controls provided by a kit (Alien SpotReport$^{\text{TM}}$ Alien$^{\text{TM}}$ cDNA Array Validation System).

Then, images were generated with the help of a laser fluorescent scanner (Agilent

G2565BA) using two wavelengths (570 nm for the green label and 660 nm for the red one). Several images per microarray were taken with different photomultiplier gains. The complete microarray experiment was performed by Dr. Oleg Tchernitsa and Technical Assistants Jana Keil and Anita Geflitter in the Laboratory of Molecular Tumor Pathology (Charité).

1.3 Image Analysis and Detection of Differentially Expressed Genes

The analysis described in the following section was done by Dr. Ralph-Juergen Kuban (Laboratory of Functional Genome Analysis, Charité).
First, spot intensities were quantified using Imagene version 3.0. If spots with saturated intensity were detected during this process, the quantification was repeated using one of the images with lower photomultiplier gain. The intensity of the local background was also quantified and subtracted from each spot.
Next, red and green intensities for each microarray pair were adjusted with the help of an MA-plot. MA-plots visualize a potential bias towards one color in a dye swap experiment by plotting the logarithmic red versus green ratio (M) dependent on the logarithmic mean intensity (A). The regression curve was then calculated with LOWESS (locally weighted polynomial regression), a nonlinear regression technique commonly applied to complex data. LOWESS is based on the fit of a low-degree polynomial to each point in the data set. It requires the specification of a certain percentage of neighbor points, the so-called smoothing parameter, which was set to 20 % in the current study. The regression polynomial is calculated locally on the neighboring points, which are weighted according to their proximity to the point in question. Finally, their value under the local regression function is calculated for all points and returned as regression curve.

After adjustment of red and green intensities, a two-way ANOVA (analysis of variance) was performed to determine the significance of differential gene expression. The application of ANOVA models to microarray analysis was introduced by Kerr et al. [KERR ET AL. 00]. They consider a four-way ANOVA model, which takes into account the influence of the array, dye, variety and gene as well as array-gene and variety-gene interactions on the (logarithmic) intensity and derive gene expression ratios from their model.
The two-way approach used here is implemented in the GeneSpring ® software version 6.1 (Silicon Genetics, Redwood City, CA). P-values were calculated using Student's t-test, assuming equal variance and Gaussian distribution of the data. Then, with the help of two-way ANOVA, the influence of dye, variety (in this case: time point) and their interaction on gene expression was estimated and a gene list re-

trieved accordingly. As multiple test correction Benjamini-Hochberg was applied [BENJAMINI & HOCHBERG 95]. Finally, a list of 82 genes, regarded as significantly differentially expressed, was derived (see Appendix, Table A).

The data set on which this work is based consists therefore of a 82 x 15 matrix, containing ratios for 82 genes over 15 time points.

2 Verification of Microarray Data

Because a microarray experiment is error-prone, the verification of results by another method is essential. In this chapter, the work of several researchers is summarized, who contributed to the verification of gene expression.

2.1 Methods

Two independent methods were used for verification: Northern blot and real-time PCR. Both are based on the same cell line that was used for the microarray experiment. The cells were treated in the same way (IPTG added at time point zero and removed after 120 h) as described in the introduction (section 1.2).

2.1.1 Real-time PCR

Real-time PCR was performed by Birte Müller [MÜLLER 04]. She verified the expression values of cpg21, Lox and Mob-1 with the house-keeping gene HPRT (hypoxanthin-phospho-ribosyltransferase) as a control.
To obtain the gene expression ratios, the difference between the limit cycle number of the control and the gene is calculated. The limit cycle number is defined as the number of PCR cycles sufficient to reach a given concentration of cDNA. It is assumed that during PCR the cDNA concentration is doubled with each cycle. Thus, the amount of cDNA increases exponentially to the base of two. The expression ratio is therefore calculated as two to the power of the limit cycle difference dC_T between gene and control: $expression(gene)/expression(control) = 2^{-dC_T}$.
A large limit cycle number corresponds to a small cDNA concentration. To account for this inverted relation between cycle number and cDNA concentration, the negative limit cycle difference is used in the equation.

Expression ratios derived from microarray and real-time PCR cannot be compared without caution, since the ratio of gene expression in the microarrray experiment is not calculated against HPRT but against the same gene under non-stimulated conditions.

2.1.2 Northern Blots

Northern blot experiments were performed by Jana Keil and Karen Weisshaupt with GAPDH as a control. Some Northern blots differ from the microarray experiment in the length of the covered time period, but the time points included in the Northern blots are the same as in the microarray experiment. In this chapter, the Northern blot experiments are classified as follows:
A
These Northern blots were done by Jana Keil. They do not include all time points from the microarray experiment, but the following: 0, 10 min, 30 min, 60 min, 2 h, 6 h, 12 h, 24 h, 48 h and 72 h. The mRNA was taken from a stock different from that used for the microarray experiment.
B
Northern blots marked with B were performed by Jana Keil and cover the same time points as the microarray experiment. In addition, the mRNA comes from the same stock that was used for the microarray experiment.
C
The third Northern blot experiment was performed by Karen Weisshaupt and covers the following time points: 0, 10 min, 30 min, 60 min, 2 h, 6 h, 12 h, 24 h, 48 h, 72 h and 96 h.

Northern blots can be semi-quantitatively analyzed, but the limited detection range of the films might lead to saturation for high concentrations of radioactive material during autoradiography. Therefore, only a rough comparison of signals derived from the control gene GAPDH and a specific gene is possible. In addition, the ratio is calculated differently from the microarray experiment (not against non-stimulated cells but against GAPDH). Another severe problem is the variation of the control itself shown in Figure 2.1.

2.1.3 Image Quantification

Quantification of Blots with ImageJ

ImageJ is a freely available software specialized on image quantification. Analysis of blots with ImageJ was performed as proposed on the ImageJ homepage (http://rsb.info.nih.gov/nih-image/manual/tech.html#analyze).
First, Northern blots were scanned using CanonScan N670U and saved as TIF files in pixel values. Next, the pixel values were transformed into optical density (OD) values with the help of a calibration curve. Ideally, this calibration curve should be derived from known mRNA concentrations. This would allow calculation of concentrations directly from the pixel values. Because these concentrations were not available, a step tablet provided by the ImageJ homepage

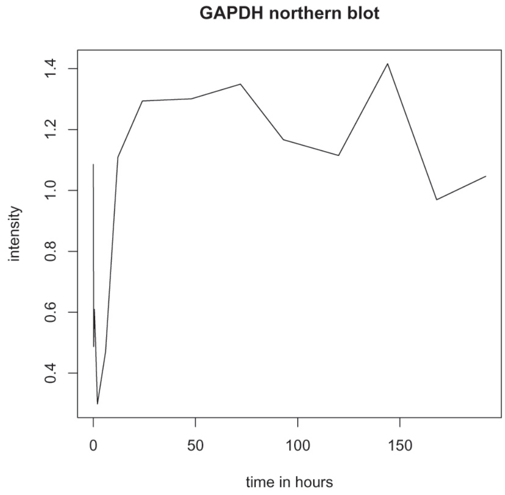

Figure 2.1: This figure shows intensity values over time for GAPDH derived from a Northern blot. For convenience, the data points are connected by a line. GAPDH was used as a control for the Northern blot experiments.

(http://rsb.info.nih.gov/ij/docs/examples/calibration/) was used to obtain a number of pixel values with known OD values. The calibration curve was then derived by fitting a Rodbard function to these known OD values.

In the next step, the background was subtracted and the OD for the lanes quantified. For this, a box was specified such that the area of quantification was the same for all lanes. The intensity calculated for each lane is the sum of the OD values over the chosen area.

Quantification of Blots with a Densitometer

As a control, the quantification was repeated with the GS-670 Imaging Densitometer for clusterin and Fra-1. Northern blots were scanned with the densitometer and analyzed with the accompanying software molecular analyst. The resulting intensity values were compared with those obtained from ImageJ.

2.1.4 Visualization and Similarity

Gene expression similarity was visualized with the freely available statistical package R (http://www.r-project.org). The similarity of two gene vectors x and y was assessed by calculating the Spearman correlation coefficient r_{sp}, which is defined as:

$$r_{sp} = 1 - \frac{6 \sum d_i^2}{(n^2 - 1)n} \tag{2.1}$$

with d_i denoting the rank differences between two elements x_i and y_i of the gene expression vectors to be compared and n the number of elements in these vectors. Because the Spearman correlation coefficient compares the ranks (the order of the values) instead of the values themselves, it is less affected by different scales.

P-values associated with the Spearman coefficients state whether the hypothesis of a slope of zero has to be accepted (no linear relationship between the expression values of two genes) or rejected (linear relationship) for a given level of significance. If for example the level of significance α is set to 0.05, a correlation with an associated p-value larger than 0.05 is not regarded as significant.

2.2 Results

In Table 2.1 the Spearman correlation coefficients of Northern blot and real-time PCR data to the microarray data are given. P-values associated with the Spearman coefficients are shown in parentheses.

Gene	Northern A	Northern B	Northern C	Real-time PCR
cpg21	0.71 (0.026)	0.77 (0.001)	-	0.63 (0.03)
Lox	0.47 (0.17)	0.59 (0.024)	0.47 (0.14)	0.43 (0.17)
Lox-related	-	-	0.66 (0.03)	-
MKP1	-	-	0.39 (0.23)	-
MKP3	-	-	0.63 (0.042)	-
Mob-1	0.47 (0.17)	0.8 (0.0006)	-	0.77 (0.005)
thrombospondin-1	-	-	0.18 (0.59)	-
Timp2 (mRNA1)	-	0.66 (0.0086)	0.55 (0.084)	-
Timp2 (mRNA2)	-	0.72 (0.0036)	-0.27 (0.42)	-
Tsc36	-	-	0.72 (0.016)	-

Table 2.1: Data from Northern blot experiments A-C and from the real-time PCR experiment are compared to those of the microarray experiment with the help of the Spearman correlation coefficient. The p-value for the correlation is given in parentheses. For Northern blots covering a shorter time period than the microarray experiment, time points not included in the Northern blot were omitted from the analysis.

If the time courses of the genes are plotted together (Figures 2.2-2.6), it can be seen that for Lox (Figure 2.2), Mob-1, MKP1 (both Figure 2.3) and thrombospondin-1 (Figure 2.4) the curve derived from the microarray data is similar to the others but shifted to the right along the time axis, resulting in low correlation coefficients. The expression vectors were standardized as described in section 3.1.2 to display gene vectors in one plot. This procedure does not affect the Spearman correlation coefficient as it does not change the ranks.

In the table below, the Spearman correlation coefficients for intensity values obtained from ImageJ and the densitometer are compared.

Gene	Spearman correlation
Fra-1	0.69 (0.006)
Clusterin	0.98 (0)

In average Spearman correlation for both methods is 0.84. Thus, intensity values retrieved by using ImageJ are comparable to those obtained by a professional densitometer.

Because of its relevance for chapter 4, the similarity between expression values of SRF obtained from a Northern blot (done by Jana Keil) and junB (microarray) was assessed (Figure 2.7). The Spearman correlation coefficient of both gene vectors amounts to 0.69 (p-value = 0.006).

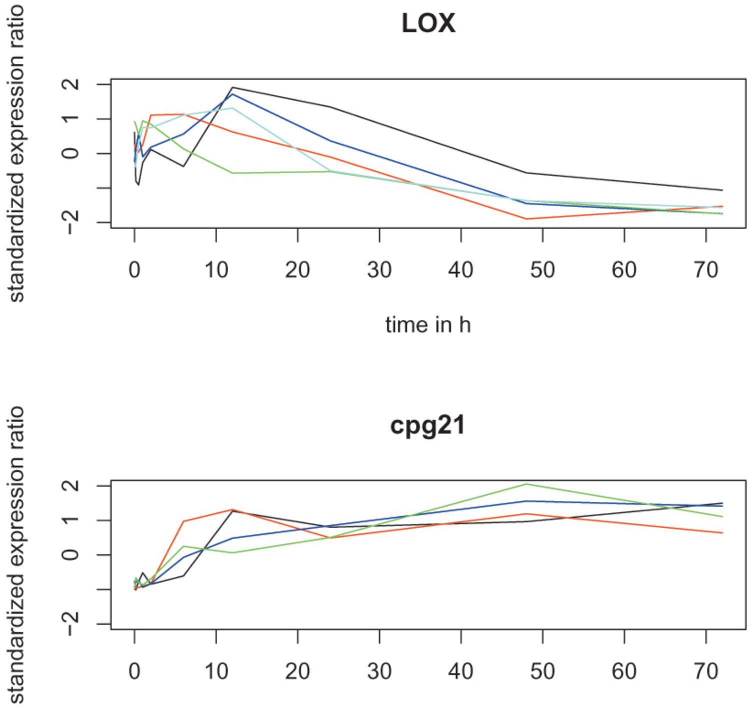

Figure 2.2: The standardized expression vectors derived from different experiments are plotted together over time. Data points are connected by a line for better comparison. The colors encode the different experiments: black = microarray, red = Northern A, blue = Northern B, cyan = Northern C, green = real-time PCR.

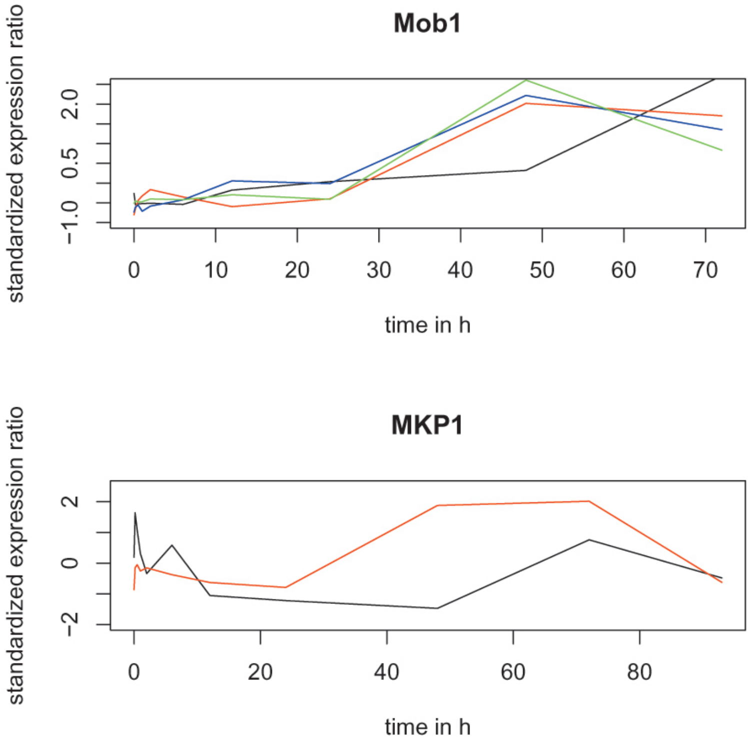

Figure 2.3: The standardized expression vectors derived from different experiments are plotted together over time. Data points are connected by a line for better comparison. The colors encode the different experiments. Mob1: black = microarray, red = Northern A, blue = Northern B, green = real-time PCR, MKP1: black = microarray, red = Northern C.

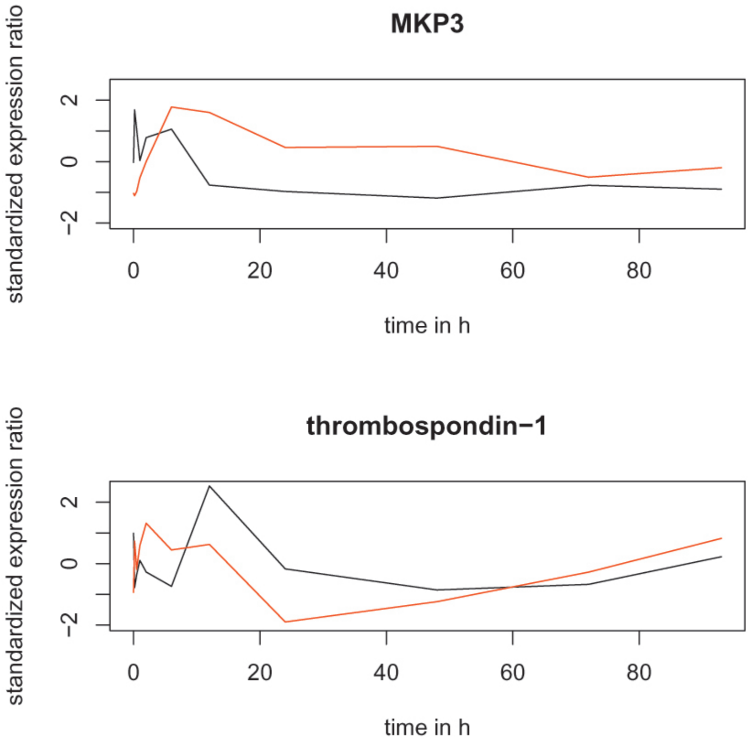

Figure 2.4: The standardized expression vectors derived from different experiments are plotted together over time. Data points are connected by a line for better comparison. The colors encode the different experiments: black = microarray, red = Northern C.

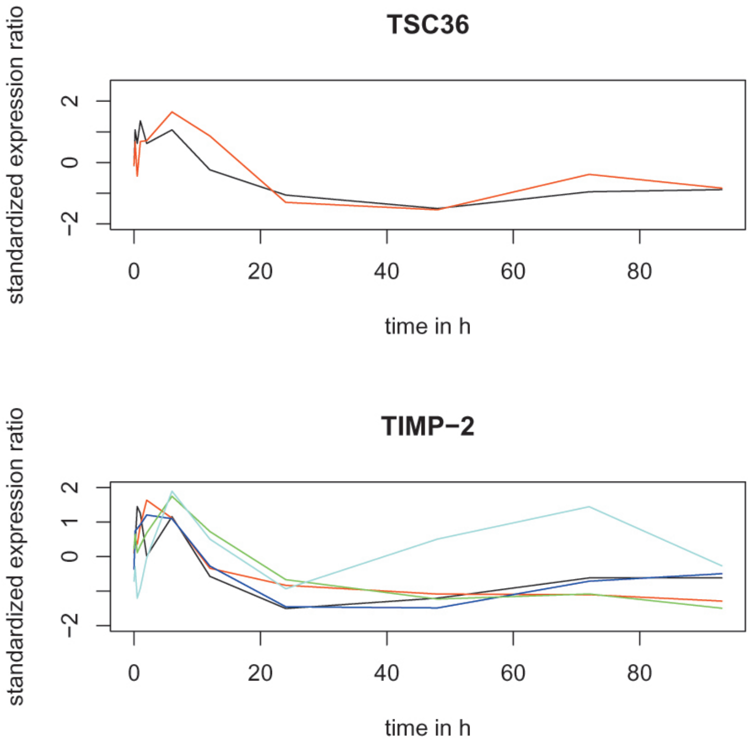

Figure 2.5: The standardized expression vectors derived from different experiments are plotted together over time. Data points are connected by a line for better comparison. The colors encode the different experiments. Tsc36: black = microarray, red = Northern C, Timp2: black = microarray, red = Northern B (1. mRNA), blue = Northern B (2. mRNA), green = Northern C (1. mRNA), cyan = Northern C (2. mRNA).

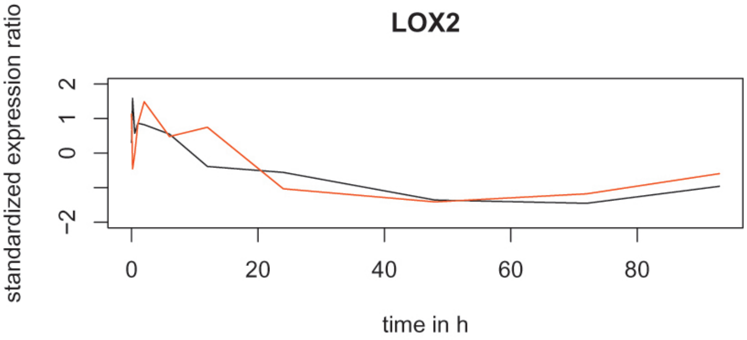

Figure 2.6: The standardized expression vectors derived from different experiments are plotted together over time. Data points are connected by a line for better comparison. The colors encode the different experiments: black = microarray, red = Northern C.

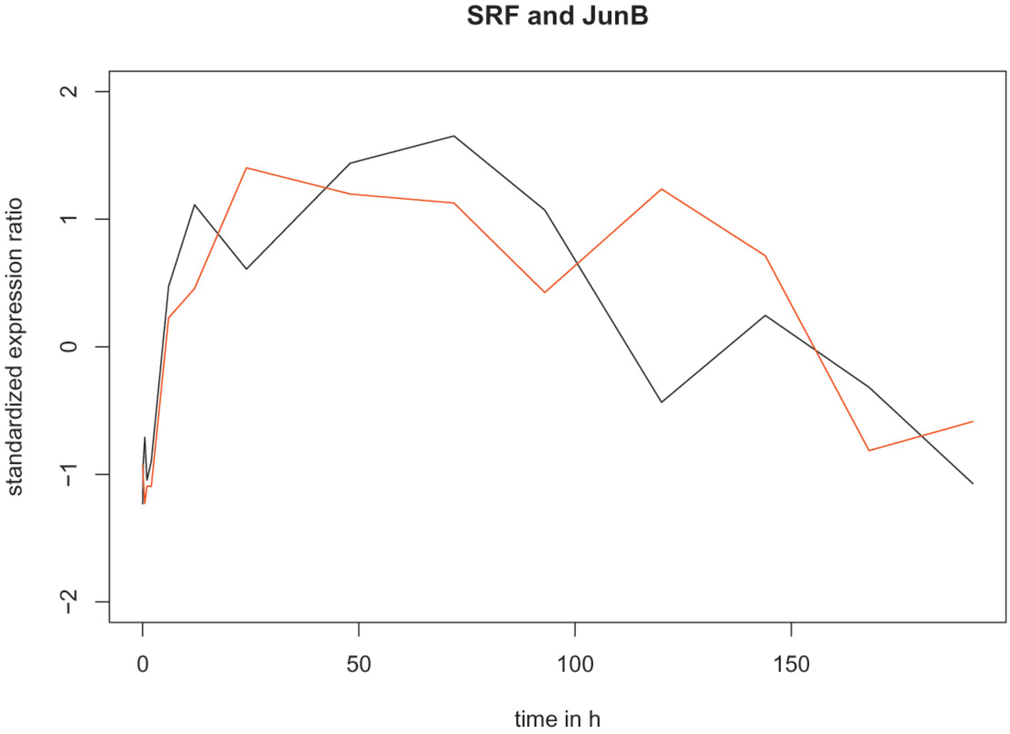

Figure 2.7: Standardized gene expression vectors of junB (red) and SRF (black) are plotted together. The data points are connected by a line for better comparison.

3 Cluster Analysis

Cluster analysis consists of several methods that find groups of similar data points in a data set. Groups should be defined such that data points inside a group or cluster are more similar to each other than to the data points outside of the group. Because cluster analysis belongs to the unsupervised data mining techniques, it requires no prior knowledge of categories in the data set and therefore is often used in microarray analysis. The basic idea of its application to microarray data is that co-expressed genes can be detected by calculating the similarity of their expression vectors and ordering the genes accordingly [EISEN ET AL. 98]. In this study, two different clustering approaches have been applied to the data set: hierarchical and partitional clustering. In addition, several methods have been used to determine the optimal cluster number k.

3.1 Materials and Methods

3.1.1 Software

The calculations described in this section were performed with R. For SOM clustering, GeneCluster 2.0 from the Cancer Genomics Group at the Whitehead/MIT Center for Genome Research was used.

3.1.2 Data Set

The data set, also called the gene expression matrix, consists of 82 genes whose expression ratios were measured at 15 time points. Therefore, one row of the expression matrix represents an expression vector for one gene. In the context of partitional clustering, a gene vector is also referred to as a data point in a 15 dimensional space.

Mean Expression Ratios

For each gene, hybridization of its cDNA with the target was performed on five different spots. Therefore, the expression ratios represent the average out of five measurements. Because the deviations from the mean expression ratios were small (negative deviation, averaged over the 82 genes: 0.09, positive deviation, averaged over the 82 genes: 0.11), only the mean expression ratios were taken into account for cluster analysis.

Logarithm

Often, the logarithm to base two is applied to microarray data to give up- and down-regulated expression ratios equal weight. In this work, no logarithm was taken prior to standardization, but its effect on clustering results was tested.

Standardization

The aim of cluster analysis is to find groups of co-expressed genes. Thus, the interest is focused on similar behavior of genes over time rather than on the magnitude of gene expression. If their vectors point into the same direction, the genes are co-expressed. Therefore, the lengths of gene vectors (magnitude of expression) can be adjusted without loss of important information, if the proportion of vector entries (determining the direction) is preserved. This is done by standardization of the data. In a procedure commonly applied, the mean of the corresponding gene expression vector is subtracted from each entry in the matrix and each entry is divided by the standard deviation of the gene expression vector. The resulting vector has mean zero and standard deviation one.

3.1.3 Positive and Negative Controls

To assess and compare the power of different clustering techniques, positive and negative controls are necessary. As a positive control, a data set derived from *S. cerevisiae* [CHO ET AL. 98] has been chosen, which consists of 416 genes manually separated in five clusters according to different phases of the cell cycle (early G1, late G1, G2, S and M). This data set has been described as suitable positive control by Futschik and Kasabov [FUTSCHIK & KASABOV 02]. The positive control was also standardized.

For generation of a negative control, each row of the gene expression matrix was shuffled such that correlations of gene vectors were destroyed. This is demonstrated by the histogram of the 82 x 82 correlations of gene vectors, which shows a Gaussian distribution centered on zero (see Figure 3.13).

3.1.4 Distance Measures

Every clustering technique requires a distance measure or metric. If a distance measure has been defined, the (symmetrical) dissimilarity matrix can be calculated whose entries correspond to the distance between two gene vectors x and y. Euclidean metric is widely used as a distance measure:

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{3.1}$$

A variety of other distance measures is mentioned in the literature. Because the focus of this project was not on the optimization of parameters but on the exploration of a given data set, Euclidean as the most popular metric has been chosen throughout this chapter. As part of their analysis of microarrays from serum stimulated fibroblasts, Herzel and colleagues tested the influence of different metrics on the clustering result and found only minor differences [HERZEL ET AL. 02]. They conclude that the choice of a certain metric is less important for clustering results.

3.1.5 Hierarchical Clustering and Heat Map

In hierarchical clustering a hierarchy of gene clusters is constructed from the dissimilarity matrix such that starting from a single cluster containing all genes the cluster number increases until each cluster consists of one gene. This results in a tree with the root (whole gene set) at the top and the leaves (single genes) at the bottom, where the branch height reflects the distance between clusters. Such a tree is also called a dendrogram [EISEN ET AL. 98].

There are two approaches for the construction of a dendrogram: one can either start from single genes and join them step by step until finally all genes are assembled in one cluster (bottom up or agglomerative strategy) or one starts from the whole gene set, splitting it in smaller and smaller clusters until the single gene level has been reached (top down strategy).

Hierarchical clustering methods can be classified according to their definition of the cluster distance (see for example [AMARATUNGA & CABRERA 04]):

1. The cluster distance can be defined as the largest distance between two data points in cluster A and B (complete linkage).

2. It can also be defined as the shortest distance between two members of cluster A and B (single linkage).

3. The average distance of all pairs of data points in cluster A and B can be used as cluster distance (average linkage).

There are additional definitions (Ward's clustering, centroid clustering), which are not considered in this work.

The R package cluster provides the function agnes (agglomerative nesting) for hierarchical clustering. As the name suggests, agnes follows the bottom up approach and allows the use of the three cluster distance definitions mentioned above. The cluster package is available at CRAN (http://cran.r-project.org/).

To assess the quality of the hierarchical clustering, the agglomerative coefficient (AC), which was introduced by Kaufman and Rousseeuw, is implemented in agnes. For each data point i, a quality measure ac(i) is specified as the ratio between the distance of i and the first cluster it joins and the distance of the cluster containing i and the last cluster it joins. The quality for the overall clustering is then defined as: $AC = \frac{1}{n}\sum_{i=1}^{n}[1 - ac(i)]$, with n as the number of data points. If AC approaches its maximum of 1, there is a high amount of clustering structure in the dendrogram.

In a heat map, the rows of the gene expression matrix are rearranged according to a given dendrogram that is displayed at the left side of the matrix. This visualization technique was first applied to microarray analysis by Eisen and colleagues [EISEN ET AL. 98]. Expression values are encoded by a range of colors. Because similar gene vectors are placed next to each other and numbers are replaced by colors, the heat map facilitates the detection of gene groups. This data representation technique is available in R with the function heatmap.

3.1.6 PCA and Partitional Clustering

The 82 x 15 gene expression matrix is a multidimensional data set consisting of 82 data points in a 15 dimensional space. Obviously, a dimension reduction technique has to be used for the visualization of clustering of the data points. Principal component analysis (PCA) is the classical method for dimension reduction and commonly applied to microarray data.

In PCA, first the underlying principal components of the data set are identified. In a second step, the projection of the data set onto the subspace spanned by the principal components is performed. In order to find the principal components, a transformation matrix representing them is searched that maximizes the variance of the data matrix X. It can be proven that the eigenvectors of the correlation matrix of X meet this requirement. Therefore, the transformation matrix is composed of the eigenvectors.
The eigenvectors associated with the largest eigenvalues explain most part of the variance in the data set and should be chosen for the projection. For this choice, a barplot of the eigenvalues, also called a Scree plot, is helpful. Projection is then performed by multiplying the chosen eigenvectors with the data matrix.
With this visualization method at hand, partitional clustering methods can be applied.

Partitional clustering methods are based on the iterative assignment of data points to a specified number of cluster centers. The assignment step is repeated until an optimum has been reached or a specified number of iterations has been executed. The definition of the cluster center and the updating step can differ, but for all these methods it is necessary to specify the cluster number k.

In this work, out of the large number of partitional cluster methods available three widely used algorithms have been chosen: kmeans, partitioning around medoids (PAM) and self-organizing maps (SOM). Kmeans and PAM are quite similar, whereas SOM is based on a different approach and therefore provides a good control.

An overview of the features of these algorithms is given in Table 3.1 below.

Algorithm	Kmeans	PAM	SOM
authors	Mac-Queen, 1967	Kaufman and Rousseeuw, 1990	Teuvo Kohonen, 1995
parameters	-k -number of iterations -metric	-k -number of iterations -metric	-grid size (m x n = k) -number of iterations -initialization of weight vectors -initial radius of neighborhood -update of radius size -initial learning rate -update of learning rate
cluster center	centroid (mean of all cluster members)	medoid (central data point)	winning node
initialization	k data points randomly chosen as centroids (R implementation)	k data points randomly chosen as medoids	random numbers or k data points randomly chosen as weights for the nodes

Algorithm	Kmeans	PAM	SOM
iteration step	-calculation of new centroids as the mean of distances of cluster members to current centroids -assignment of data points to nearest centroid	-calculation of cluster cost for all data points -calculation of cluster cost for all data points after swapping each data point with its medoid -new medoid: data point which reduces cluster cost most	for each randomly chosen input data point: -identification of winning node (node with weight vector most similar to input vector) -adjustment of weight vector of winning node to input vector according to learning rate -adjustment of weight vectors of neighboring nodes either in a distance-dependent or all-or-nothing manner
termination	change of centroids below threshold or given number of iterations completed	change of cluster cost below threshold or given number of iterations completed	given number of iterations completed
advantages	fast (linear order)	less sensitive to outliers than kmeans	-both robust and accurate -well suited for large data sets
disadvantages	-sensitive to outliers -possibly different results on re-run	-time consuming for larger data sets	-large number of parameters to specify -possibly different results on re-run

Table 3.1: The information on kmeans and PAM was taken from [AMARATUNGA & CABRERA 04], SOM and its application to microarrays is described in [TAMAYO ET AL. 99].

3.1.7 Cluster Validation

Partitional clustering methods require the specification of the cluster number k. This parameter can be determined by several strategies:

Data Visualization

Careful inspection of the heat map gives a first insight into similar behavior of genes. Also, the distribution of branch heights in the dendrogram might point to a certain cluster number.

Quality of Clustering

Another approach for determining k is the assessment of clustering quality in order to find an optimum for a certain cluster number. A variety of quality measures has been introduced, from which the Davis-Bouldin index, Dunn's index and the silhouette index have been chosen. In the work published by Bolshakova et al. the application of all three indices to microarray data is described [BOLSHAKOVA & AZUAJE 03]. The Davis-Bouldin index was calculated as given in [GÜNTER & BUNKE 02].

Davis-Bouldin index

First, the similarity between two clusters A and B is defined:

$$d(A,B) = \frac{\sigma_A + \sigma_B}{d(c_A, c_B)} \tag{3.2}$$

whith c_A and c_B denoting the cluster centers of clusters A and B, whereas σ_A and σ_B are defined as the average distances of cluster members in cluster A and B to their centers. Then, Davis-Bouldin index is given as follows:

$$DB = \frac{1}{k} \sum_{A=1}^{k} \max_{B=1..k, B \neq A} d(A,B) \tag{3.3}$$

For a good clustering (large distances between cluster centers) DB should approach its minimum (zero).

Dunn's index

Dunn's index is defined as:

$$Dunn = \frac{d_{min}}{d_{max}} \tag{3.4}$$

where d_{min} is the shortest distance between a pair of data points in two different clusters and d_{max} denotes the largest distance between two data points in one cluster. For a good clustering, d_{min} should be much larger than d_{max}. Therefore, the larger Dunn's index the better the quality of the clustering.

Silhouette index

The silhouette index was introduced by Kaufman and Rousseeuw. For a single data point it is given by:

$$sil(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{3.5}$$

where $b(i)$ denotes the average distance of a data point i to all data points in the nearest neighbor cluster, whereas $a(i)$ describes the average distance of data point i to all data points in the same cluster. For a good clustering $b(i)$ is much larger than $a(i)$ and $sil(i)$ approaches its maximum of one. The silhouette index for the overall clustering is defined as the average silhouette index over all data points.

Variability of Cluster Assignment

It is desirable to assess the variability of cluster memberships to know how reliable clusters are. Pollard and van der Laan propose the application of nonparametric bootstrap for this task [POLLARD & VAN DER LAAN 05].

The basic idea of nonparametric bootstrap is to simulate bootstrap data sets by drawing with replacement from the original data set and to calculate the distribution of the parameter in question over the artificial data sets. As it is nonparametric, no assumptions about the distribution are made.

In case of cluster memberships the distribution of the label vector θ is of interest. The label vector describes an assignment of genes to clusters, so the vector may consist of colors or numbers from 1 to k. Then, an observed clustering result (a labeling) is defined as $\theta_I = S(X_I)$, where I is the sample size, X_I is the observed gene expression matrix and S is a certain rule applied to the data matrix, that is a cluster method. θ and X, the true cluster result derived from the true data set, are not known (notation follows [POLLARD & VAN DER LAAN 05]).

Now, the re-sampling vector is generated by drawing with replacement from the number of columns of X. As Efron pointed out ([EFRON 92]), this results in a multinomial distribution of re-sampling vectors where the probability for the re-sampling vector that reproduces the original sample is higher than the probability for any other re-sampling vector. The columns of the data matrix are rearranged according to the re-sampling vector and thus, the new bootstrap sample is obtained.

For each bootstrap data set X_I^*, the labeling vector θ_I^* is calculated by reassigning the genes to the medoid that is now closest to them. The bootplot function developed by Pollard et al. plots each gene as a bar, colored according to the cluster membership of the gene. Thus, if a gene had been assigned to a red cluster in half of the bootstrap samples and to a green one in the other half, its bar would be colored half red and half green. The number of times a gene appears in a given cluster is called reappearance proportion by the authors. The authors recommend a number of 1000 bootstrap samples. They also point out that calculating bootstrap memberships is a form of fuzzy clustering.

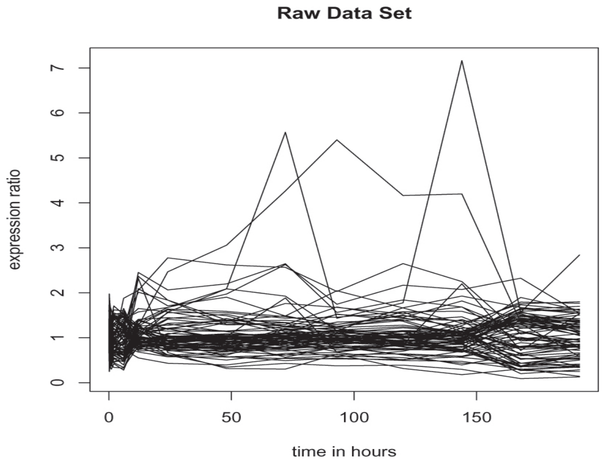

Figure 3.1: Raw data set: The expression ratios of 82 genes are shown over 15 time points. Each gene is repesented by a line connecting the time points. The large single peak belongs to MMP10, whereas the vector with the two large peaks represents Mob1.

3.2 Results

3.2.1 Data Set

The raw data set is dominated by the expression of Mob1 and MMP10 (see Figure 3.1). The Mob1 expression peak at 144 h represents the highest up-regulation (about 7 fold) observed in the data set. Fibronectin was the gene most strongly down-regulated (about 10 fold).

In the standardized data set (Figures 3.2 and 3.3), two groups of genes can be recognized: one that is up-regulated between 12 and 144 h and another one that is down-regulated in the same period. Interestingly, there are two sudden changes. The first change is visible between 6 and 12 h, the second between 144 and 168 h.

3.2.2 Hierarchical Clustering

Agglomerative clustering with average linkage was used as hierarchical clustering technique. Choosing another linkage method leads to changes in the overall structure of the dendrogram. But in the heat map (see Figure 3.4) the same four gene groups are visible, although different linkage methods were used.

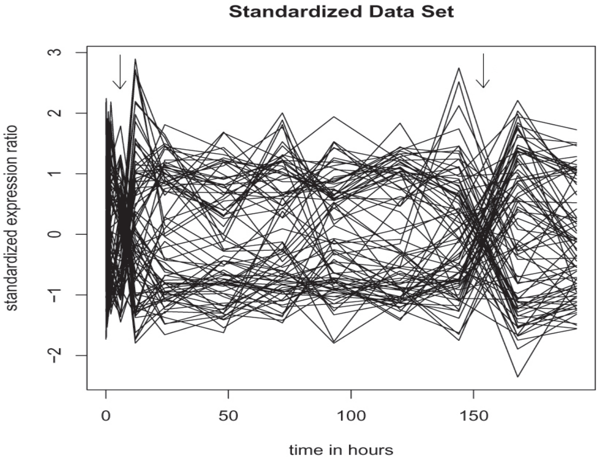

Figure 3.2: Standardized data set: The expression ratios were standardized as described in the text. For the 82 genes, the time-dependent expression ratios, connected by a line, are shown. The two changes between 6 to 12 h and 144 to 168 h are indicated by arrows.

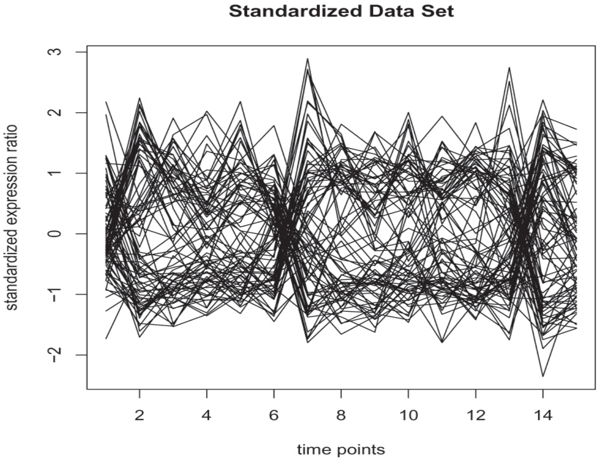

Figure 3.3: Standardized data set: Here, the expression values of the 82 genes are not shown on the time scale. Instead, each of the 15 time points is displayed together with its expression ratios, which are connected by a line. Thus, the first 6 expression ratios of the genes are more clearly visible than in Figure 3.2.

These four groups can be described as follows:

1. The first large group contains genes that are down-regulated from 0 to 6 h, than up-regulated and again down-regulated after 144 h, that is 24 h after IPTG removal (group B).

2. The second large group shows the opposite behavior: its genes are up-regulated from 0 to 6 h, than down-regulated and again up-regulated after IPTG removal (group A).

3. There is also a small group of genes that peaks at 12 h and is down-regulated after IPTG removal (group D).

4. The last group is down-regulated from 0 to 6 h and stays up-regulated after this time point (group C).

The table below gives ACs for the three linkage methods (positive and negative control clustered using average linkage).

Method	Average	Complete	Single	Negative control	Positive control
AC	0.74	0.76	0.63	0.37	0.7

The AC points to a detectable structure in the data set. The comparatively high AC for the negative control can be explained by the fact that for each gene the expression values at given time points are swapped, but not altered. Thus, even in the shuffled data set there is some structure present: the genes which are in average up-regulated and those in average down-regulated.

3.2.3 PCA

The first and the second principal component together explain 77.3 % of the variance in the data set. In addition, there is a sharp decline after the first two eigenvalues visible in the Scree plot (see Figure 3.5). Therefore, only the first two components are used for the analysis.

The genes are mainly separated according to their expression values between 6 and 144 h. Genes on the right side of the biplot (see Figure 3.6) are up-regulated in this time period, those on the left side down-regulated.

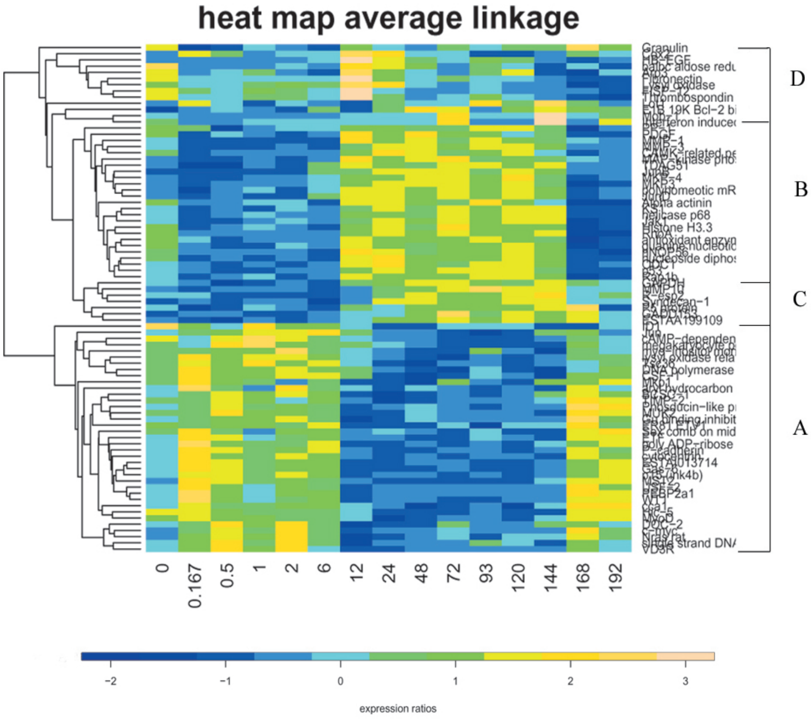

Figure 3.4: Here, the heat map resulting from average linkage is shown. On the left side of the gene expression matrix, the dendrogram is displayed. The expression ratios are encoded by colors ranging from blue (down-regulated) to orange (up-regulated). On the right side, a rough assignment of the genes to clusters A-D is shown.

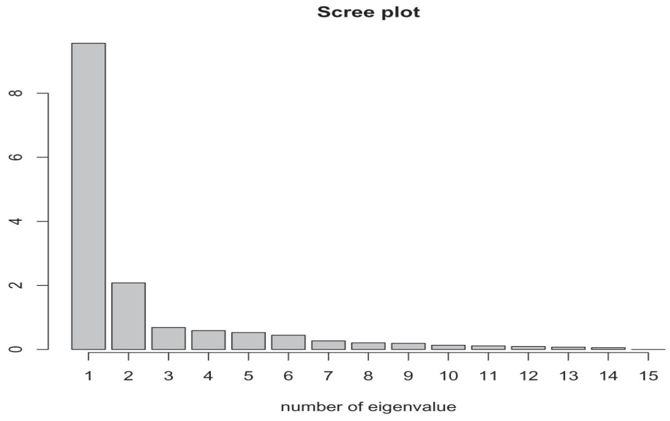

Figure 3.5: Screeplot of the eigenvalues. A sharp decline is visible after the second eigenvalue.

3.2.4 Partitional Clustering

All three cluster procedures were performed with k equals 4 and the Euclidean metric. The issue of choosing k will be discussed in depth in the next section. SOM was run with the default values given by the authors of GeneCluster:

Parameter	Value	Explanation
grid size	2 x 2	defines the shape of the SOM
neighborhood	bubble	sharp (all-or nothing) decline of node vector adjustments around the winning node
number of iterations	50,000	
initial radius of neighborhood	5	the radius covers the nodes whose weight vectors are adjusted during the update step
initial learning rate	0.1	adjustment of weight vectors of neighbor nodes
final radius of neighborhood	0.5	radius of neighborhood decreases in each step
final learning rate	0.005	learning rate decreases in each step
initialization of node weights	random vectors	

The output of the clustering procedures, the label vector, was used to color the genes in the biplot according to their cluster memberships. Such a coloring of the 82 genes is shown in Figure 3.6. It is of note that the blue cluster in Figure 3.6 corresponds to the genes with a peak at 12 h (group D), whereas the black cluster represents the genes mainly down-regulated (group A) and the green cluster those mainly up-regulated (group B). The red cluster contains those genes which are up-regulated but not down-regulated upon IPTG-removal (cluster C).

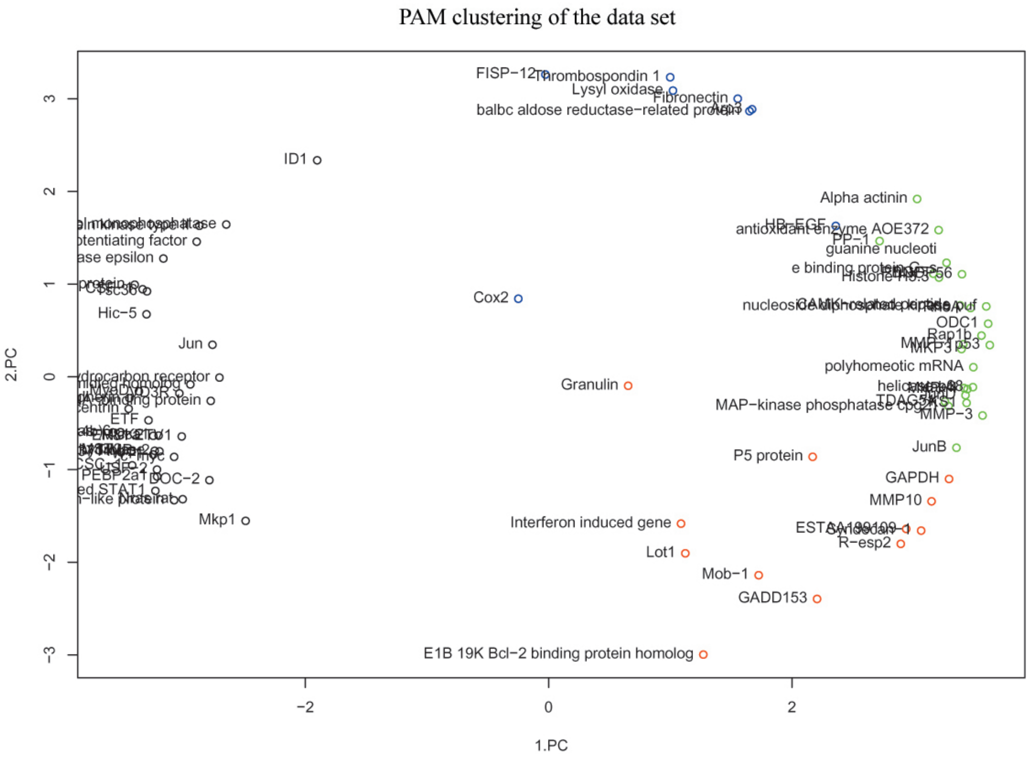

Figure 3.6: Data points representing the 82 genes are plotted into the space spanned by the first and second principal component. The points are colored according to the label vector resulting from PAM clustering.

If the label vectors resulting from the three algorithms are compared, it can be seen that they assign the genes to the same clusters (only kmeans leads to a deviating label vector for some runs). Therefore, the clusters identified can be regarded as stable.
The profiles of the four clusters identified are given in Figure 3.7. They are displayed together using different colors in Figure 3.8.

Gene lists of the clusters together with their gene ontology (GO) terms can be found in the Appendix (Tables B.1-B.4).

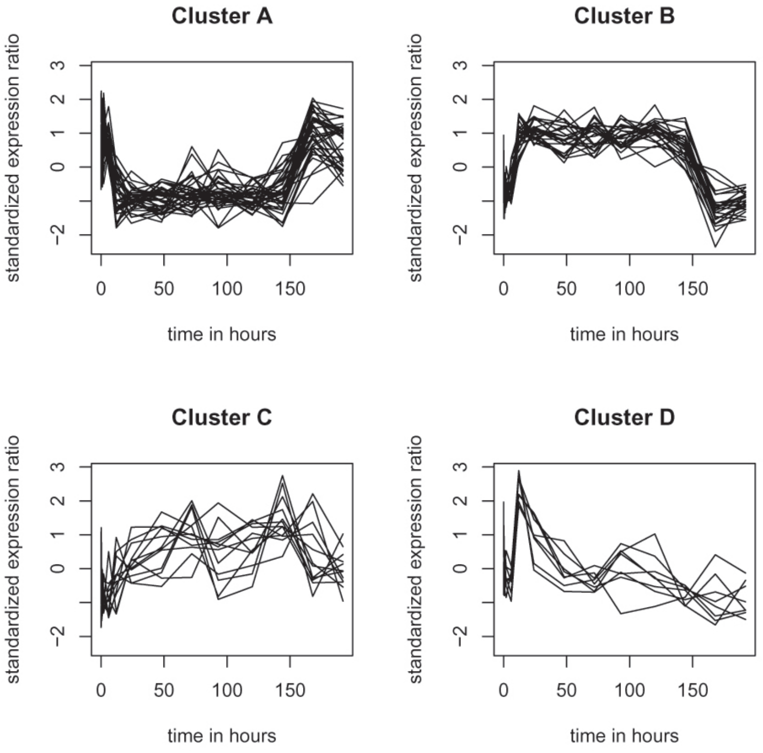

Figure 3.7: Four clusteres were identified in the data set using kmeans, PAM and SOM. Here, the time-dependent expression ratios of the genes belonging to one cluster are displayed for clusters A-D. For convenience, these values are connected by a line.

3.2.5 Controls

PAM was used to cluster the controls, as it is the most reliable of the three cluster algorithms. Again, k was set to four.

To check whether taking the logarithm prior to standardization would affect the results, the logarithmic data set was standardized and clustered. Taking the logarithm did not change the partitioning vector for the given cluster number.

Clustering was also done on the positive (with k set to five) and the negative control. Figure 3.9 displays the 416 genes of the positive control in a biplot. Their cluster membership according to the label vector generated by PAM is encoded by five

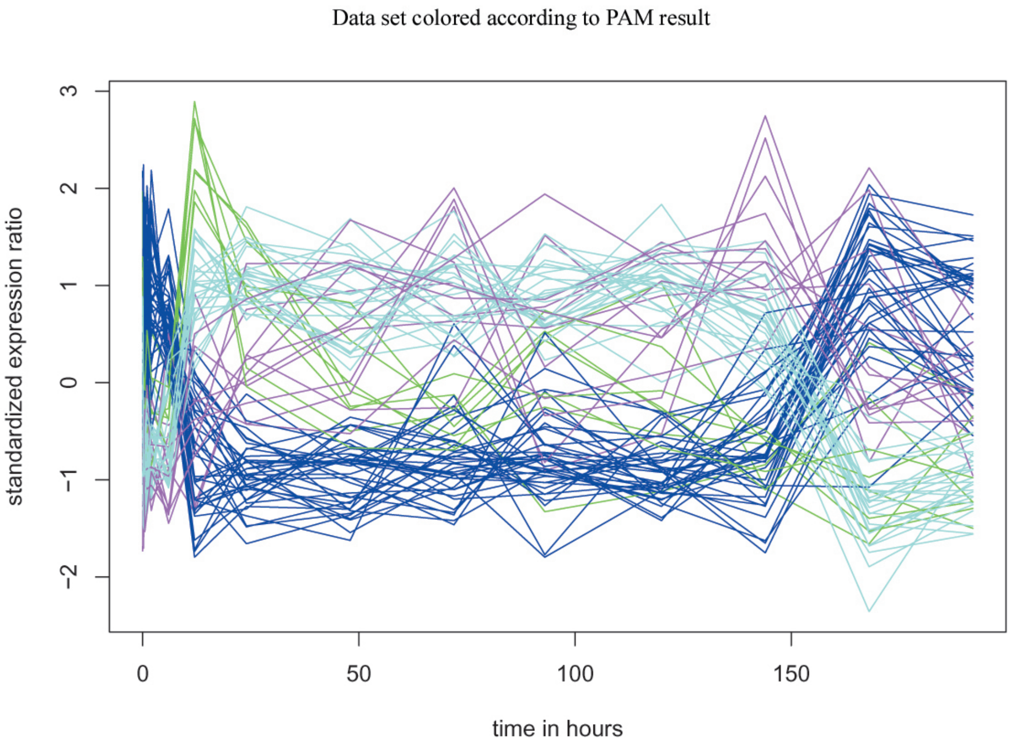

Figure 3.8: Gene expression vectors of the data set colored according to their cluster memberships as calculated by PAM. Blue corresponds to cluster A, cyan to cluster B, magenta to cluster C and green to cluster D.

different colors. It can be seen that the red and the blue cluster overlap.
If the time course of each gene in the positive control is colored according to the cluster membership assigned manually (Figure 3.10) and according to the label vector generated by PAM (Figure 3.11) a difference can be seen especially for those genes peaking between 0 and 50 h. However, the profiles of the clusters identified by PAM are similar to those found by manual assignment.
The clustering of the negative control is shown in Figure 3.12. It can be seen that a certain structure is present that reflects overall up- or down-regulation of a gene. Figure 3.13 shows that correlations of gene vectors near to zero are most frequent in the shuffled data set.

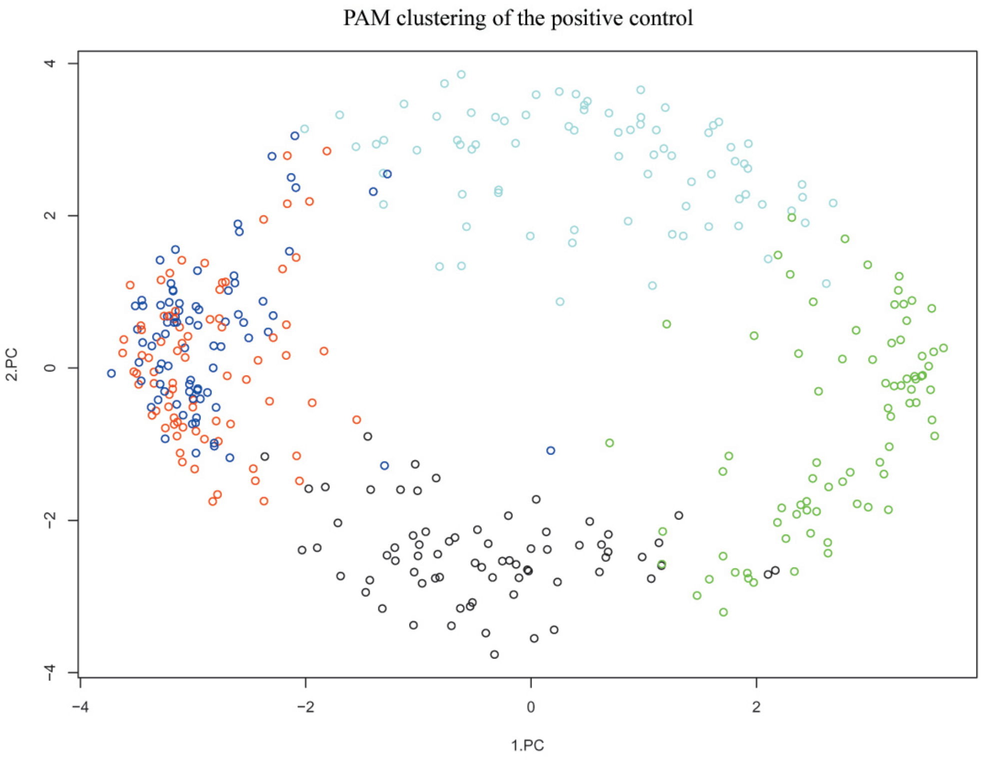

Figure 3.9: The 416 gene vectors forming the positive control are displayed as data points in a biplot and are colored according to their cluster membership as calculated by PAM. It is of note that the red and the blue cluster overlap.

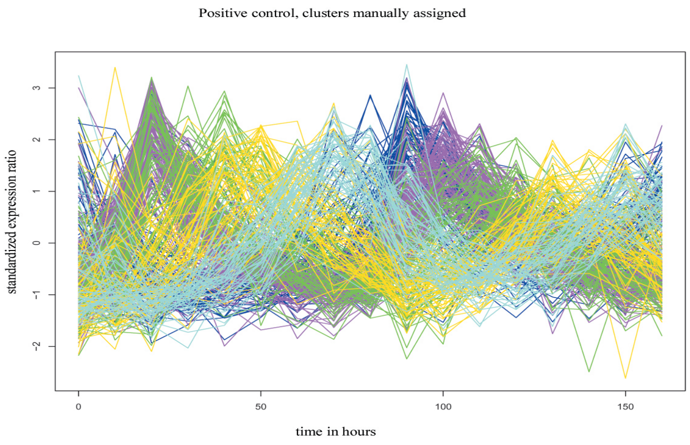

Figure 3.10: Gene expression vectors of the positive control colored according to manually assigned clusters.

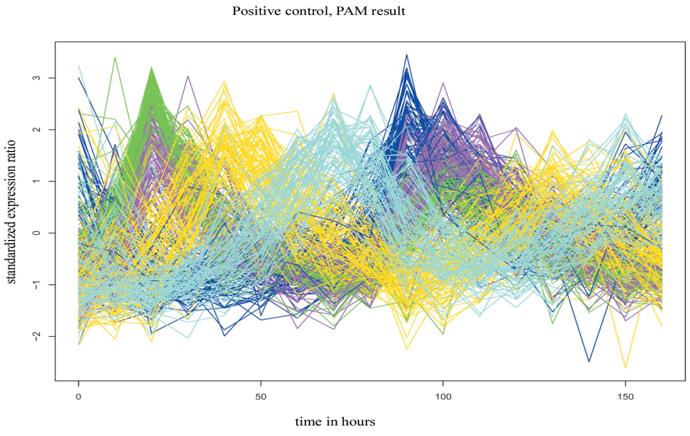

Figure 3.11: Gene expression vectors of the positive control colored according to the label vector generated by PAM.

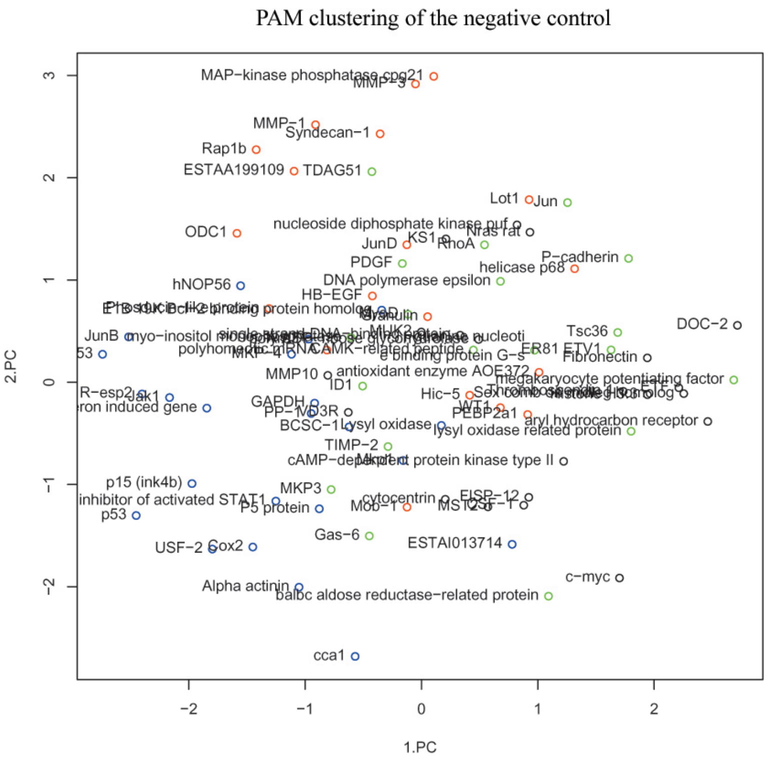

Figure 3.12: The row-wise shuffled gene expression matrix was used as negative control. Here, the shuffled gene vectors are shown in a biplot and colored according to the label vector generated by PAM.

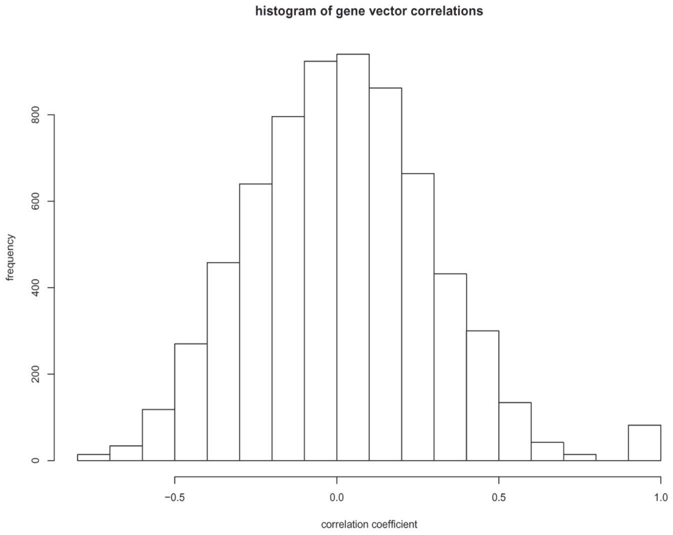

Figure 3.13: Here, a histogram of Pearson correlations of gene expression vectors from the shuffled data set is shown. The bar at a correlation coefficient = 1 represents the diagonal of the correlation matrix.

Another question was the influence of the first 6 time points on the cluster result. In chapter 2 it was observed that for some genes (Lox, MKP1, Mob-1, thrombospondin-1) expression values from the microarray experiment are shifted in comparison to values derived from Northern blots and real-time PCR. Therefore, it is of interest whether skipping the first 6 time points (0, 10 min, 30 min, 60 min, 2 h, 6 h) affects cluster results.

Without the first 6 time points, the blue (cluster A) and the magenta cluster (cluster C) become poorly separable (Figure 3.14). If the first 6 time points are replaced by their mean and clustering is repeated, the result is comparable to the one obtained by skipping the first 6 values (see Figure 3.15). Thus, the mean of the first six expression ratios does not give additional information for cluster separation in the modified data set.

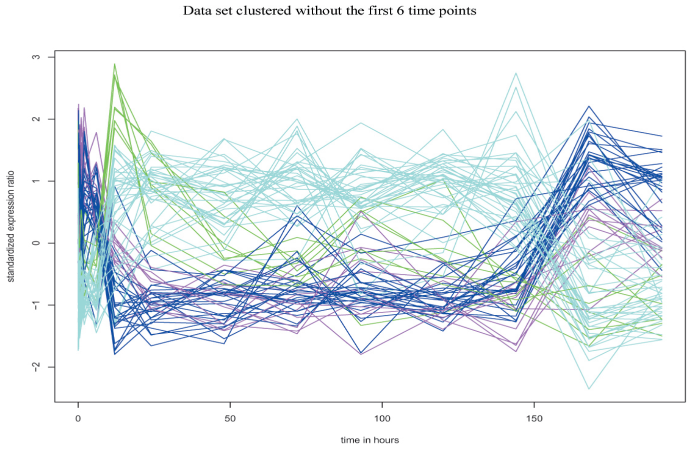

Figure 3.14: Gene expression vectors are colored according to cluster memberships that were calculated by skipping the first 6 time points.

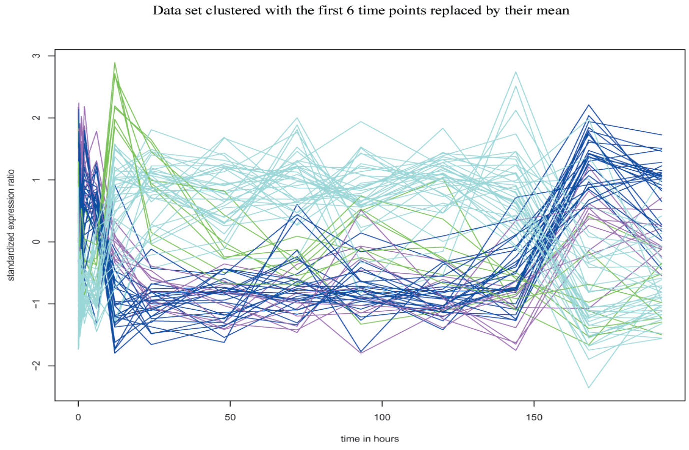

Figure 3.15: Gene expression vectors are colored according to cluster memberships that were calculated by replacing the first 6 time points with their mean.

3.2.6 Choosing k

Visualization

All three heat maps show the same four gene groups. The dendrogram derived from the dissimilarity matrix using average linkage has two prominent branches, pointing to a cluster number two. The branch heights of the other dendrograms are evenly distributed and therefore less easy to interpret (data not shown).

Quality of Clustering

For three indices (Davis-Bouldin, Dunn's index and silhouette index), the quality of clustering was calculated for k in a range between 2 and 80 (see Figure 3.16).

It can bee seen that Davis-Bouldin and Dunn' index are biased, because for the random data set cluster quality increases with cluster number k (see Figure 3.17). Therefore, only the values for k between 2 and 20 were further investigated. A larger cluster number is not regarded as biologically meaningful for this small gene set. In the table below, the k indicated as optimal by the cluster indices is given for a range of k between 2 and 20. The values of the indices for the optimal k are added in parentheses.

Index	Davis-Bouldin	Dunn's	Silhouette
optimal k	2 (0.087)	2 (0.53)	2 (0.52)

If the indices are applied to the positive control in the same range, they fail to detect 5 clusters as the optimum (see Figure 3.19). Nevertheless, the values of all three indices decrease abruptly for k $\geq$ 5. Interestingly, the data set also shows a decline of cluster quality for k larger 5 (Figure 3.18). This points to a cluster number not higher than 5, which fits well the pattern seen in the heat map.

Bootstrap

The bootplot for a cluster number of four reveals well defined clusters with high average cluster memberships (Figure 3.20). It is of note that some clusters are better defined than others, for example cluster number zero has a higher average cluster membership than cluster number 1. Thus, the average cluster number reflects the quality of each cluster.

The question arises whether the average cluster membership can be used for the optimization of k. To address this question, the cluster memberships are averaged over all clusters. Then, this average membership index is calculated over a range of k by using 1000 bootstrap data sets for each k. As this is computationally very intensive, the range for k was restricted from 2 to 20. The average membership index also points to a cluster number not larger than 5 (see Figure 3.21).

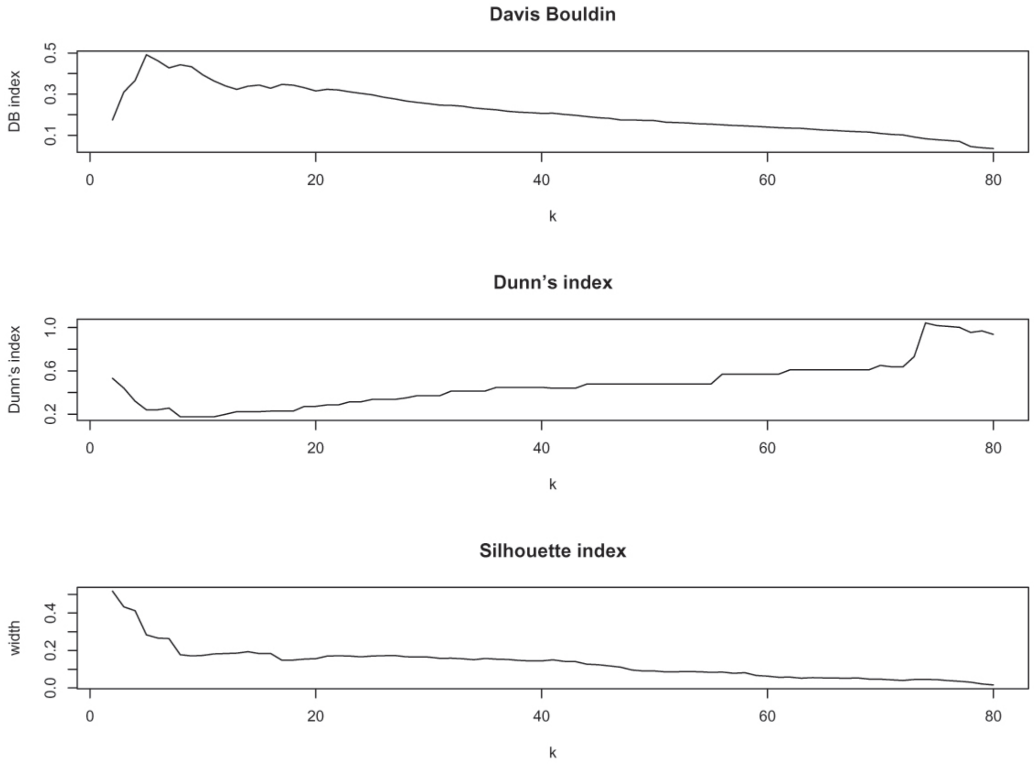

Figure 3.16: The values of the three indices Davis-Bouldin, Dunn's and silhouette for a k ranging from 2 to 80 are shown (values connected by a line). Davis-Bouldin and Dunn's index point to an optimum of 80 clusters in the data set.

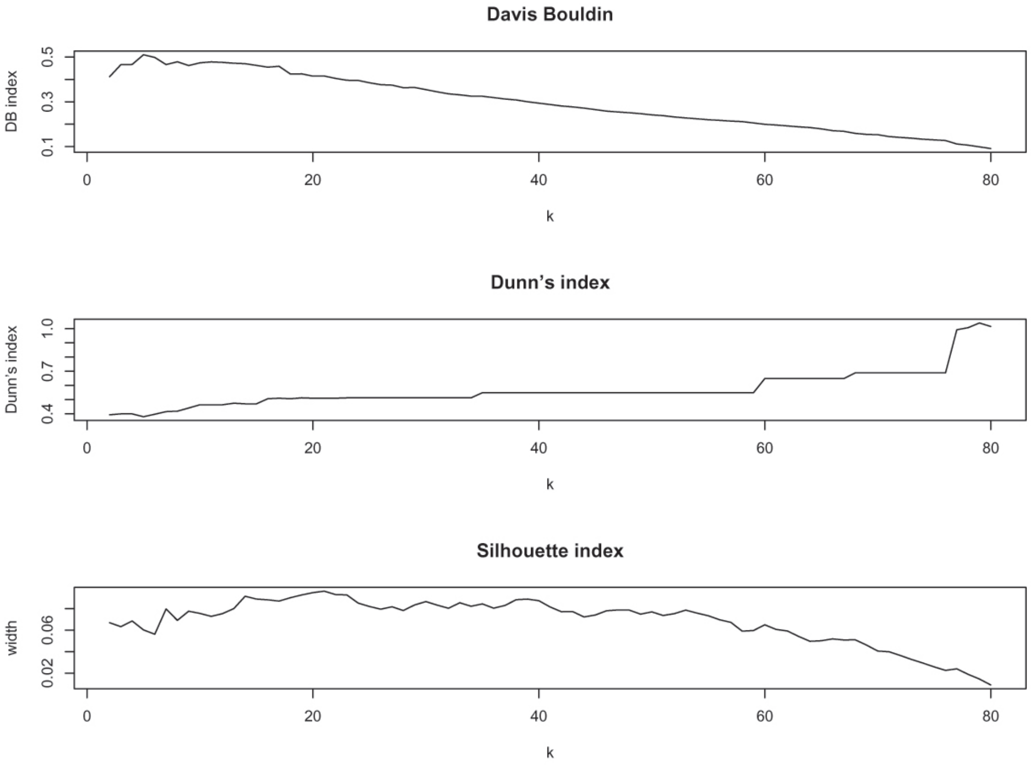

Figure 3.17: The values of the three indices Davis-Bouldin, Dunn's and silhouette are given for the negative control. As in Figure 3.16, Davis-Bouldin and Dunn's index point to an optimum of 80 clusters. The values of the indices for k from 2 to 80 are connected by a line.

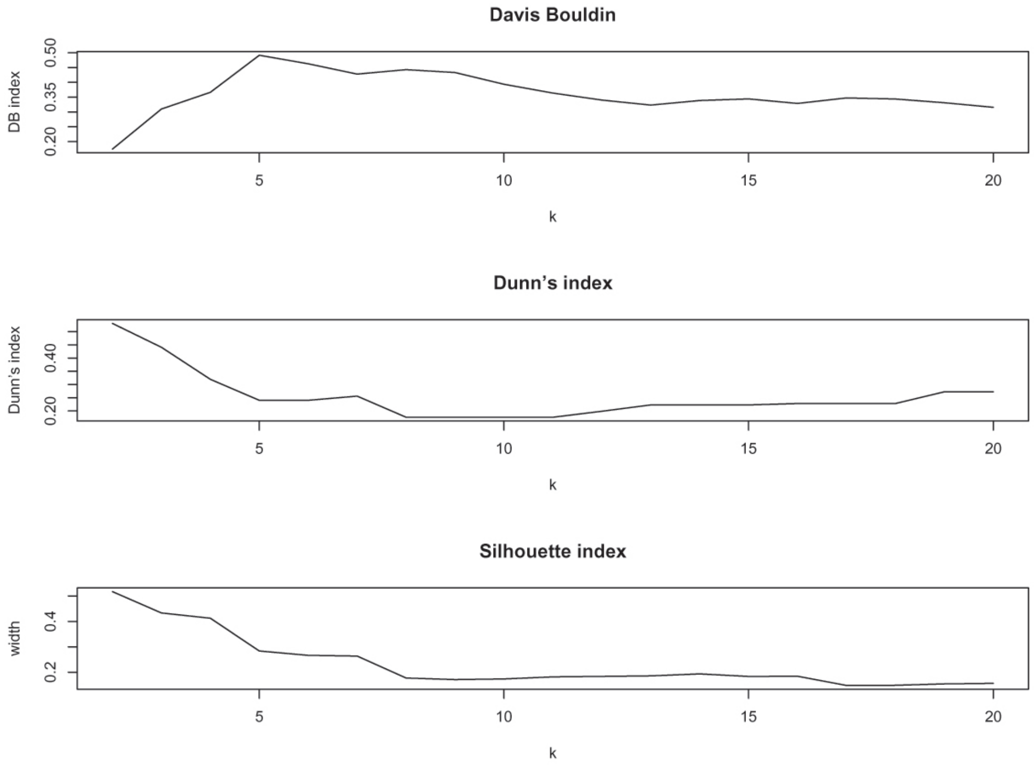

Figure 3.18: The values for the three indices Davis-Bouldin, Dunn's and silhouette are given for the data set for k ranging from 2 to 20.

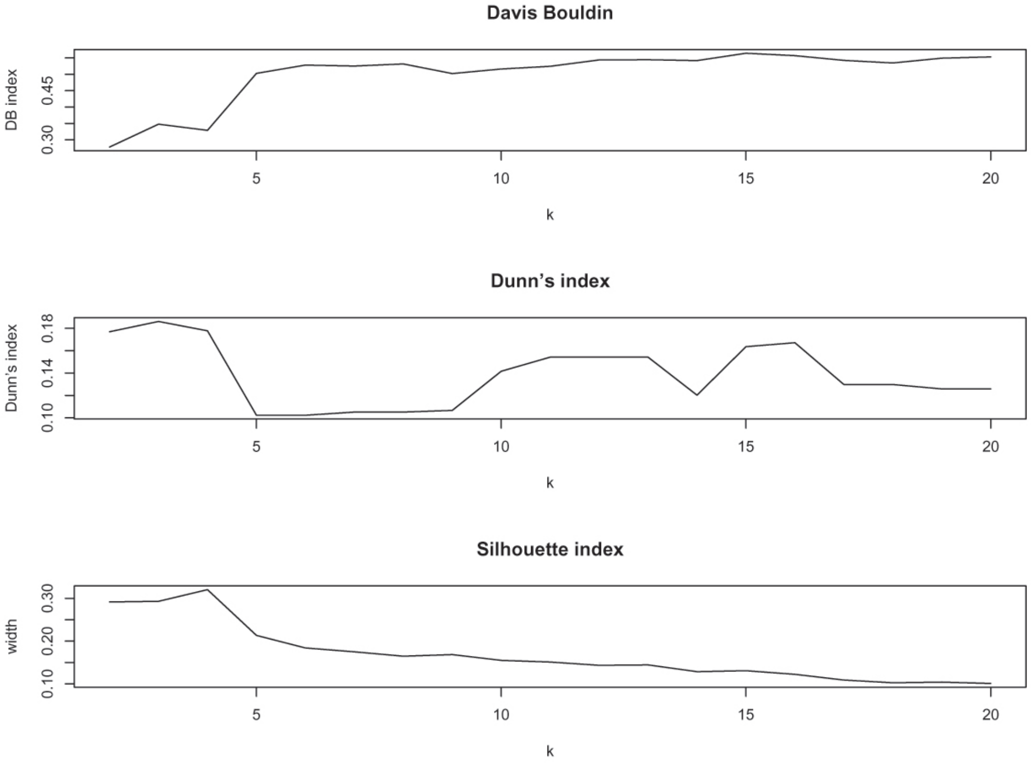

Figure 3.19: The values of Davis-Bouldin, Dunn's and silhouette are displayed for the positive control for k ranging from 2 to 20.

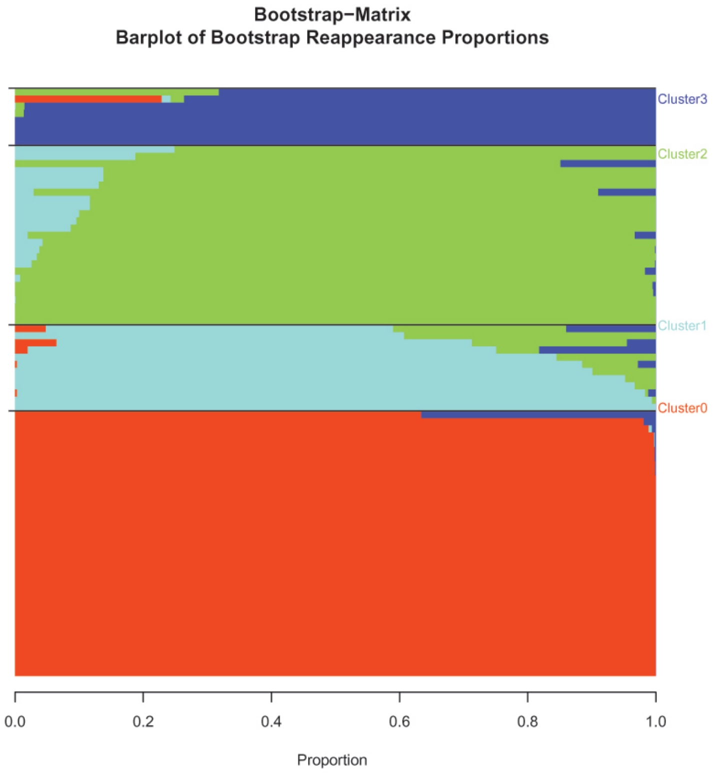

Figure 3.20: Bootplot of the cluster memberships for k equals 4. Each bar represents one gene and is colored according to the clusters this gene belongs to. The reappearance proportion refers to the frequency a gene appeared in the same cluster in the simulated data sets. Number zero corresponds to cluster A, one to cluster C, two to cluster B and three to cluster D.

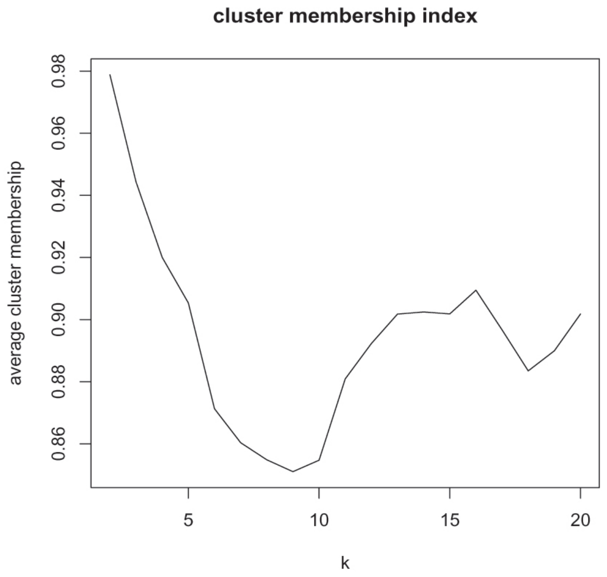

Figure 3.21: The average cluster membership values are plotted dependent on k with k ranging from 2 to 20. For convenience, they are connected by a line.

3.3 Discussion

Four clusters have been identified in the data set using a variety of cluster methods. It has to be noted that the ratios of the gene expression matrix are rather weak, but gene expression vectors correlate well with vectors derived from verification experiments.

Comparison with other Microarray Experiments

In the following, the data set under investigation is compared with three data sets derived from similar microarray experiments.

Iyer and colleagues measured 517 differentially regulated genes in serum stimulated human fibroblasts over a time span of 24 h and separated them in 10 clusters [IYER ET AL. 99]. The authors report the down-regulation of MKP-1 within 6 hours and, more interestingly, a down-regulation, followed by an up-regulation after 6 h for junB. In addition, Cox2 is found to be up-regulated after 16 h. Their findings agree well with the data derived from the current experiment.

The experiment by Tullai et al. shows up-regulation of MKP-3, cpg21 and junB in human glioblastoma cells stimulated with PDGF [TULLAI ET AL. 04]. In contrast to the results of the microarray experiment presented here, MKP-1 was found to be highly up-regulated.

Another microarray study, recently performed on normal versus Ras-transformed embryonic mouse fibroblasts, confirms down-regulation of thrombospondin-1 and Lox in the transformed cell line [VASSEUR ET AL. 03].

Problematic Aspects of Cluster Analysis

A severe problem is the pre-selection of genes according to their significance. This rules out all weakly regulated genes. Thus, any cluster method is bound to find at least two clusters of regulated genes: up- and down-regulated genes. However, the peak at 12 h seen in cluster D is hard to explain by pre-selection, since removing weakly regulated genes does not cause the formation of peaks among highly regulated genes. Thus, the four clusters identified do not only represent the result of a pre-selection.

As Figures 3.11 and 3.10 show, there are differences between the clusters found by PAM and those derived manually for the positive control. In Figure 3.10 it can be seen that manually assigned groups of gene expression vectors overlap. This demonstrates one of the limits of the cluster methods used: they fail to detect the correct clusters if these overlap.

Another shortcoming of the cluster methods applied in this section is the assignment of genes to only one cluster. This might not reflect the biological situation, where genes can be regulated by more than one pathway. In contrast to kmeans, fuzzy c-means assigns to a gene not an integer (zero/one) but a percentage as membership value for a given cluster, thus allowing a gene to be a member of several clusters. The benefits of fuzzy c-means for the analysis of microarray data have been described recently ([FUTSCHIK & KASABOV 02]). Due to time constraints, fuzzy c-means was not applied to the data set.

When using the cluster quality indices on the random data set, it is striking that Dunn's index and Davis-Bouldin index are biased by the cluster number k. Both cluster indices are based on the ratio of cluster variance and cluster center distances. In the case of Davis-Bouldin, cluster variance is defined as the average of cluster member distances to the cluster center and appears in the nominator. For Dunn's index, the cluster variance is the largest distance between two members of the same cluster and stands in the denominator. With increasing k, cluster variance approximates zero, whereas cluster distance also increases. This means that Davis-Bouldin approaches zero (cluster distance in the denominator) whereas Dunn's increases to infinity (cluster distance in the nominator).
Therefore, both indices reach their optimum when most clusters consist of only one gene. In contrast to Dunn's and Davis-Bouldin, silhouette index relies not on global cluster distances but calculates distances between neighboring clusters. That might be the reason why it is less sensitive to increasing k.

Skipping of the first 6 data points results in the melting of two clusters. This demonstrates on the one hand that the first data points are important for cluster discrimination and on the other that cluster C is less stable than the others.
In Western blots performed by Jana Keil with the same cell line under the same conditions as described in section 1.2, it can be seen that Ras concentrations increase rapidly within the first 6 hours (data not shown). Thus, the Ras-dependent up- and down-regulation of many genes might be delayed by the time Ras needs to reach high concentrations after induction.
In this context Figure 3.3 is interesting, because it shows that genes in the first 6 time points (0 to 6 h) form two groups that are inversed between the sixth and the seventh time point and inversed a second time after IPTG removal. Thus, the two groups seen between the first and the sixth time point, where the concentration of the HRAS oncogene probably has not reached its maximum, correspond to the groups found between timepoints 13 to 15, where Ras-overexpression is switched off.
If the first 6 data points indeed represent a delay, gene expression ratios in this time period are not relevant for cluster analyis, since they are not influenced by Ras overexpression or reversal of transformation. In this case, there might be only three instead of four clusters in the data set.

Discussion of Cluster Members

Cluster A
This cluster is remarkable due to its large number of known tumor suppressors (Doc-2, ETF, p15-ink4b, P-cadherin, WT1, Tsc36, Gas-6, TIMP-2, MKP-1) that are down-regulated. More importantly, these genes are up-regulated again as soon as IPTG is removed, demonstrating the reversibility of their suppression. The mechanism of Ras signaling and suppression of these genes is currently unknown.

Cluster B
This cluster contains immediate early genes like junB that are rapidly up-regulated during the first 6 hours and retain an elevated expression level until IPTG-removal. A particularly high number of transcription regulating genes is found in this cluster (JunB, JunD, RhoA, Rap1b, KS-1, p53). The up-regulation of known Ras targets related to transformation (MMP1, MMP3) confirms earlier findings obtained from IR-4 cells and other cell models [SERS ET AL.02]. It is of note that cluster B also contains the dual specificity phosphatases MKP-3 and MKP-4, known to act in the cytoplasm as deactivators of P-MAPK. The over-expression of oncogenic Ras therefore enhances the negative feedback loop exerted by the DUSPs. Interestingly, the important tumor suppressor p53 is also up-regulated.

Cluster C
Cluster C is the least well defined of all four clusters and consists of genes that are up-regulated from the start but in contrast to cluster B not clearly down-regulated upon IPTG-removal. Striking in this cluster is the existence of two prominent peaks, hinting for the periodic expression of some cluster members (especially Mob-1 and interferon induced gene). The cluster contains tumor suppressors (Lot-1, GADD153) as well as genes associated with transformation (MMP10, Granulin, Syndecan-1) and the two most strongly over-expressed genes in the whole data set: Mob-1 and MMP10.

Cluster D
Cluster D is a heterogeneous group containing three known tumor suppressors (FISP-12, thrombospondin-1 and Lox), two genes involved in metabolism (Cox2, balb-c) and two others related to cell shape and motility (Arp, fibronectin). Although functionally diverse, these genes show a remarkably similar behavior with a prominent peak at 12 h, followed by down-regulation, which is not reversed upon IPTG-removal. As has been noted by O. Raudies (paper in preparation), these genes might not participate in the phenotypic reversion of Ras-transformed cells since their expression values are not affected by IPTG-removal. Interestingly, Lox can be repressed in non-induced IR-4 cells by adding medium obtained from induced IR-4 cells. Thus, a secreted factor might be responsible for Lox repression (an autocrine loop involving the EGF receptor

is already reported for Ras-transformed MCF-10A cells [SCHULZE ET AL. 01]). Blocking of the MEK/ERK-pathway with a MEK-inhibitor prevented Lox down-regulation in the medium-treated non-induced cells [SERS ET AL. 02]. This demonstrates that the MEK/ERK-pathway is involved in Lox repression. A similar blocking experiment with another cell line done by the same authors demonstrates the dependence of thrombospondin-1 down-regulation on the MAPK-cascasde.

In preceding experiments Sers and colleagues noticed that down-regulation of genes has the same importance for transformation as up-regulation [SERS ET AL. 02]. This is clearly visible in the standardized data set where roughly half the genes is up-regulated and the other half down-regulated. Another interesting point is that most clusters contain tumor suppressors as well as oncogenes. The co-regulation of oncogenes and tumor suppressors makes it hard to infer simple relationships between gene regulation and transformation. This underlines the complexity of Ras signaling and the pathways involved.

4 Screening for Serum Response Factor Targets

If genes are co-expressed, they might be regulated by the same transcription factor. Thus, clustering provides candidate groups for transcription factor binding site (TFBS) screening. Of the four clusters identified in the previous chapter, cluster B is the most promising for investigation of TFBS, because it shows a high similarity of gene expression vectors and fast up-regulation. Interestingly, this cluster contains three known serum response factor (SRF) targets and in addition the time course of gene expression is very similar to the time course of SRF observed from a Northern blot. Therefore, members of cluster B were chosen for SRF target screening.

4.1 Serum Response Factor and its Targets

Before presenting the method and discussing the results of the SRF screen, important information concerning SRF and its known binding sites is given in this section.

4.1.1 Biological Background

SRF

SRF belongs to the MADS-box family of transcription factors. The MADS-box, located at the N-terminus of the SRF protein, contains a DNA-binding domain, a dimerization domain and an interface for protein-protein interactions [MIANO 03]. The binding site of SRF, also known as the CArG-box, has the core motif CC(A/T)$_6$GG.
It is of note that this motif is nearly the same if read from the opposite strand and the opposite direction. Such a symmetry is also called a palindrome and typical for DNA sequences that are recognized by dimeric proteins. Deviations from the core motif are known for some promoters (for example SRF, CArG-box 4 [BELAGULI ET AL. 97]).
SRF binds to the CArG-box as a homodimer. The two terminal G residues of the CArG-box on both DNA strands are important for contact formation, since their mutation disrupts binding of SRF to the motif. The AT-core is also strongly preserved. Its composition of nucleotides forming only two hydrogen bonds eases the observed bending of the CArG-box [MIANO 03].

SRF targets can be separated into two functional groups: muscle related and immediate early genes [GINEITIS & TREISMAN 01]. Induction of immediate early genes (IEGs) by SRF requires a ternary complex factor (TCF), whereas induction of muscle related genes by SRF is TCF independent. Because it is often found in serum-induced IEGs, the region that combines the TCF binding site (also called the Ets-motif) and the CArG-box is known as serum response element (SRE).

Ternary Complex Factors

TCFs are a subfamily of the Ets-domain transcription factors (Ets proteins). In mammals, three TCFs are known: SAP-1, SAP-2 (Net/ERP) and Elk-1. ETS-proteins share four conserved domains: the Ets DNA-binding domain that recognizes Ets binding sites in target genes, the C- and D-domain and the B-box. The D-domain of the Ets-protein interacts with P-MAPK, which leads to the phosphorylation of the C-domain and activation of the Ets-protein. Activated TCFs have a higher ability to bind DNA and to form ternary complexes together with SRF and DNA than inactive TCFs. In these complexes, contact between SRF and TCF is mediated by the B-box [VICKERS ET AL. 04]. The core motif recognized by all Ets-proteins has the sequence GGA(A/T) [GRAVES & PETERSEN 98].

Ternary Complex Formation

Mo et al. succeeded in crystallizing the ternary complex formed by SAP-1, SRF and c-fos SRE [MO ET AL. 01]. The structure of the complex clearly showed the binding of SRF and SAP-1 at opposite sites of the DNA and bending of the DNA. Interestingly, the authors describe contact formation between SRF and nucleotides outside of the CArG-motif. They suggest that the nucleotides flanking the CArG box mediate differential binding of SRF to DNA in presence and absence of SAP-1.

The distance between the CArG-box and the Ets-motif can vary, because the B-box is attached to the TCF by a flexible linker. Also, the relative orientation of both motifs is not fixed [BUCHWALTER ET AL. 04].

4.1.2 Known and Predicted SRF Targets

Muscle-related Genes

In his review, Miano lists known, muscle-related SRF binding sites together with their distance to the transcription start site (TSS). Most of these CArG-boxes are positioned 2-3 kb around the TSS. The author suggests that this distance is of biological relevance, since interaction of RNA polymerase II with SRF is known to take place [MIANO 03].

Immediate Early Genes

The focus of the current study is on IEGs rather than muscle-related genes. For IEGs it is known that SRF regulation includes TCF formation. Therefore, an overview of genes with validated SREs is given in Table 4.1. Three of them (mcl-1, egr-1, c-fos) are described as TCF-regulated in [VICKERS ET AL. 04], junB is mentioned to be regulated by a ternary complex in [BUCHWALTER ET AL. 04]. It is of note that junB is a member of cluster B.

Gene (species)	SRE sequence (distance from TSS)	Additional CArG-boxes	Validation methods	Authors
mcl1 (human)	CCGGAAGCTG CCGCCCCTTTCC CCTTTTATGG (Blat: -61) (authors: -128)	-	luciferase reporter assays	[Townsend ET AL.98]
egr-1 (mouse)	1) GGAAACG CCATATAAGG (authors: -415) (Blat: -414) 2) CGGAACAGA CCTTATTTGG (authors: -379) (Blat: -379) 3) CCTTATATGG AGTGGAGTGG CCC$(N)_{37}$GG CTCTGGGAGGA (authors: -355) (Blat: -353)	-	site-specific mutations in combination with luciferase reporter assays	[Clarkson ET AL.99]
junB (mouse)	CTTCCTGTGC CCTAATATGG (authors:-1462) (Blat: -1740)	CCATATATGG (authors:+2084) (Blat: +1812)	promoter region: reduction of response to mitogens upon site-specific mutation downstream: luciferase reporter assays	[PHINNEY ET AL.95] downstream: [PEREZ-ALBUERNE ET AL.93]

Gene (species)	SRE sequence (distance from TSS)	Additional CArG-boxes	Validation methods	Authors
c-fos (human)	ACAGGATGT CCATATTAGG (authors: -216) (Blat: -329)	-	site-directed mutation analysis, transcription of actin gene with synthetic c-fos CArG-box, crystallization of ternary complex	sequence annotation given in [Mo et al. 01], distance to TSS and validation methods except crystallization given by Wasserman and Fickett (http://www.cbil.upenn.edu/MTIR/HomePage.html)

Table 4.1: This table lists known SREs. The first column gives the name of the gene and the organism where its promoter was characterized. The second column displays the sequence with CArG-box and Ets-motif underlined. The distance to the TSS as indicated by the authors and as detected by using BLAT is given in parentheses. A minus (-) signifies a position upstream of the TSS whereas a plus (+) stands for a position downstream of the TSS. The third column lists additional CArG-boxes that are not part of a SRE. In the fourth column, verification experiments are given and the last column lists the sources of information used.

For the following two CArG-box containing genes it is unclear whether they are regulated by SRF alone or whether TCF formation takes place. Therefore, they are not listed in Table 4.1 but discussed separately.

PDGF-A

In the human promoter of PDGF-A, a CArG-box 477 bp upstream of the TSS (sequence: CCTTTTATGG) was validated [Lin et al. 92]. An Ets-site is not mentioned in this promoter study. Like junB, PDGF-A is also a member of cluster B.

SRF

SRF is known to contain four different CArG-boxes and one Ets-site [Belaguli et al. 97] with the following sequences (numbering of the CArG-boxes according to the authors):

S4:	CCTTTAAGG
S3:	CAAATAAG
S2:	CCATATAAGG
S1:	CCATAAAAGG
Ets:	GCTGGAATT

Concerning the formation of a ternary complex at the SRF promoter, contradictory information is given in the literature. According to Spencer et al., the murine SRF gene is induced upon basic fibroblast growth factor stimulation either by the Ras pathway via Sp1 or by the RhoA/Rac1 pathway via SRF binding to the CArG-box without involvement of a ternary complex [Spencer et al. 99].
In contrast, in a more recent study on fragments of the mouse promoter, it is stated that SRF indeed is induced upon serum stimulation by a ternary complex composed of Elk-1 and SRF [Kasza et al. 05]. Inspection of the murine SRF promoter published by Belaguli et al. reveals that the Ets-site is separated by more than 150 bp from the nearest CArG-box [Belaguli et al. 97]. These different findings might be reconciled if the existence of another Ets-site in close proximity to a CArG-box is assumed. It might also be possible that TCF formation can take place over large distances.

However, both studies establish a link between Ras signaling and SRF induction and point to the existence of a positive feedback loop. Because of this loop, the gene expression curve of SRF is expected to be similar to those of its targets. Therefore, the similarity of the time course of cluster B to that of SRF indicates potential regulation of cluster members by SRF.

Experimental Evidence

Recently, SRF knockout experiments were performed using Affymetrix microarrays and a gene list of potential SRF targets was derived, whose members were further analyzed for the occurrence of CArG-boxes [Philippar et al. 04]. The known SRF targets junB, egr-1, egr-2 and cyr61 were detected as well as three new target genes, namely tuftelin-1, fhl2 and keratin-17.

Another group studied the effects of blocked pathways on gene expression with the help of microarrays [Tullai et al. 04]. Four large groups of genes were obtained: PI3K- and MEK-independent, MEK-dependent, PI3K-dependent and PI3K- and MEK-dependent. These groups were further screened for SRF binding sites within 1 kb upstream of the TSS. Thus, 16 CArG-boxes in 10 promoters were predicted, 13 of them conserved in mouse. To validate these findings, the authors used chromatin immunoprecipitation and measured the enrichment of CArG-containing promoters over GAPDH. Among the novel SRF targets identified in the MEK-dependent group is cpg21 (DUSP5), a member of cluster B. In addition, the known target genes egr-1 and fos were confirmed.

Predictions

The prediction of CArG-boxes performed by Dieterich et al. is exclusively based on phylogenetic footprinting. With the help of alignments, conserved non-coding sequence blocks (CNBs) were found. These CNBs were screened with string representations of known TFBS from TRANSFAC. Alternatively, a scan with weight matrices was done. Interestingly, junB and DUSP2 were among the predicted candidate genes [DIETERICH ET AL.03].

In a recent study presenting the CNB database CORG, it is stated that upstream regions of SRF are conserved in human and rodents as well as in fish. This points to the presence of the positive feedback loop not only in mammals but also in other vertebrates [DIETERICH ET AL.05].

4.2 Material and Methods

A summary of the prediction procedure is given in Figure 4.1.

4.2.1 Software

In general, scripts used in this section were written in Perl. Screening was done with BioMinerva, a collection of modules implemented by Steffen Grossmann based on BioPerl. BioMinerva allows easy handling and efficient storage of sequences, as it fully supports the gff format. This format implements definitions and their relationships given by the sequence ontology (SO).
The SO consists of a structured set of terms defining sequence properties that are connected via is-a (a gene is a sequence) and part-of (an exon is part of a gene) relations. Because of its well defined concepts, the SO and its representation by gff is particularly valuable for sequence annotation and was used throughout this work.

Both the SO terms and the description of the newest version of gff (gff3) are freely available under http://song.sourceforge.net/gff3.shtml.

4.2.2 Sequence Retrieval

Window Size

The length of the sequences under investigation has to be balanced carefully. On the one hand, by choosing a short sequence one might miss motifs, on the other hand, a long sequence increases the number of false positives.
In case of SRF, the summary on CArG-box distances from the TSS given by Miano

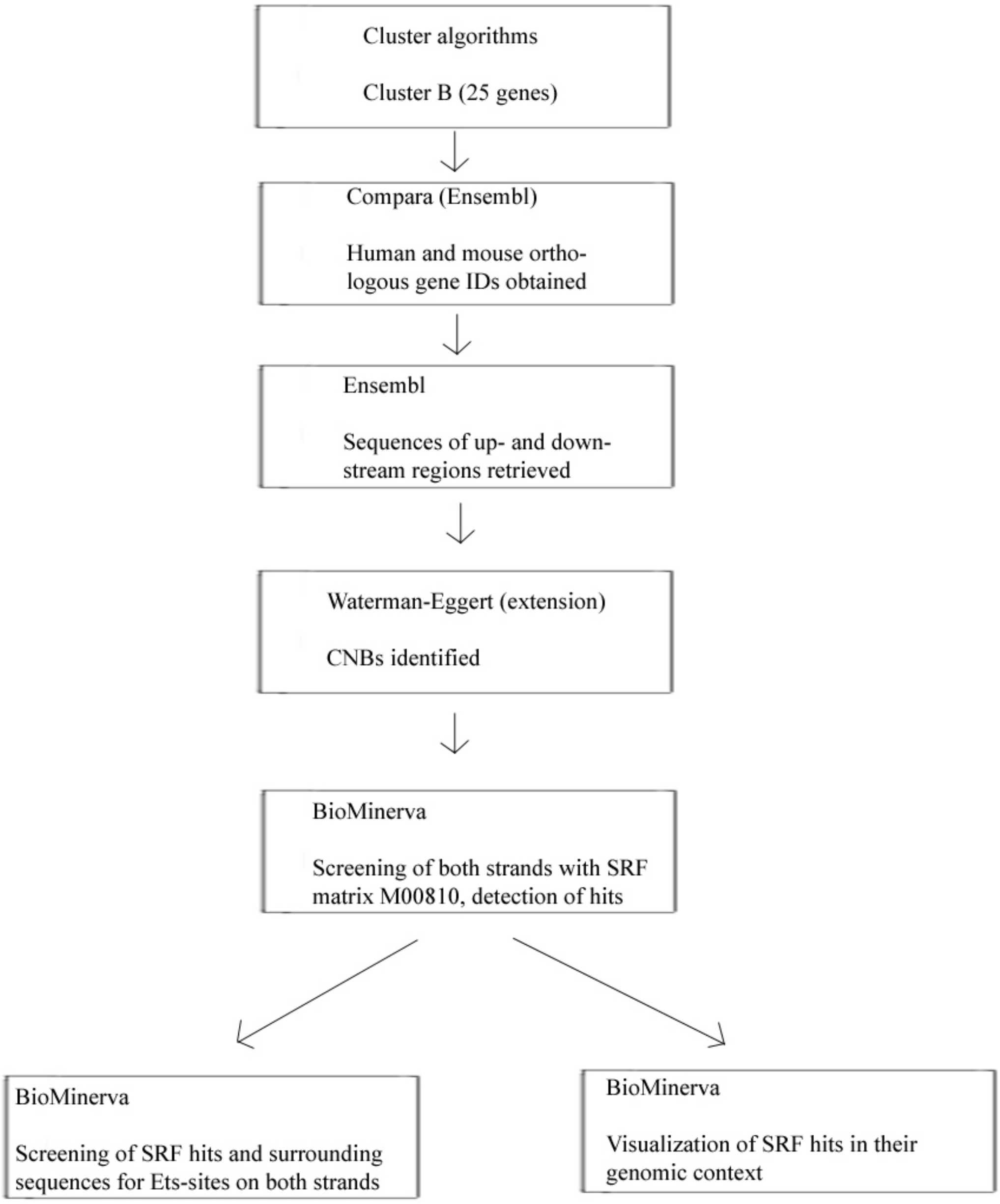
Cluster algorithms
Cluster B (25 genes)
Compara (Ensembl)
Human and mouse orthologous gene IDs obtained
Ensembl
Sequences of up- and downstream regions retrieved
Waterman-Eggert (extension)
CNBs identified
BioMinerva
Screening of both strands with SRF matrix M00810, detection of hits
BioMinerva
Screening of SRF hits and surrounding sequences for Ets-sites on both strands
BioMinerva
Visualization of SRF hits in their genomic context

Figure 4.1: This flow chart summarizes the procedure used for the prediction of SRF targets.

indicates that most known SRF binding sites are covered by a 5 kb window up- and downstream of the TSS. Therefore, a sequence window of this size was screened upstream from the TSS together with the first intronic region. If several transcripts were available, additional regions resulting from a shifted TSS were also screened. Because downstream regions are known to contain CArG-boxes (for example junB), a region 5 kb downstream of the end of transcription of the longest transcript was also retrieved. It is assumed that most transcription factor binding sites are situated upstream of the TSS or in some distance from the end of transcription. Therefore, introns (except for the first) and exons were not screened as for long genes this would increase the multiple testing problem. Due to this restriction, some TFBS might be missed.

Database

Sequences have been retrieved from the Ensembl database with the help of the Ensembl Perl API and saved in fasta format. Repeats in the sequences were masked. As gene models stored in Ensembl are updated on a two-month basis, one version was chosen for each species and used throughout this work.

Human:	homo_sapiens_core_26_35
Mouse:	mus_musculus_core_26_33b
Rat:	rattus_norvegicus_core_26_3d

Ensembl (version 26) gene predictions are based on the NCBI35 assembly for the human genome, on the RGSC v3.1 assembly for the rat genome and on the NCBI33 assembly for the mouse genome. In addition, ab initio predicted genes in Ensembl are cross-checked with cDNAs stored in RefSeq during the Ensembl prediction pipeline [CURWEN ET AL.04]. Therefore, gene models obtained from Ensembl agree well with cDNAs from RefSeq.

Validation of TSS

Because Ensembl gene predictions are not always reliable, the TSS given by Ensembl was compared with the TSS stored in the DBTSS (DataBase on Transcription Start Sites [SUZUKI ET AL.04]) and in the EPD (Eukaryotic Promoter Database [PÉRIER ET AL.00]).

The EPD comprises a non-redundant collection of experimentally verified and annotated promoters. In this database, a promoter is defined as the upstream gene region next to the TSS. The EPD is restricted to RNA polymerase II binding sites of higher eukaryotes and accessible via a web interface.

The DBTSS gives the sequence and the genomic position of full-length cDNAs (which include the 5’-end of their corresponding mRNA). The data is based on a newly developed method for the construction of full-length cDNA libraries, named oligo-capping. The newest version of DBTSS covers several species, among them mouse and human. It also provides cross-references to RefSeq and other databases.

Promoter sequences given in the EPD were mapped onto the genome with the blat program provided by the UCSC genome browser ([KAROLCHIK ET AL. 03]).

4.2.3 Phylogenetic Footprinting

The prediction of TFBS inevitably leads to a high number of false positives. To enhance specificity of prediction two techniques are frequently employed: phylogenetic footprinting and detection of combined TFBS, so called regulatory modules [WASSERMAN & SANDELIN 04].

Phylogenetic footprinting is based on the assumption that selective pressure on TFBS leads to their conservation in different species. Species have to be chosen carefully to avoid too much or too few conservation between them [WASSERMAN & SANDELIN 04]. It was suggested to compare man and mouse sequences, since these two species display highly conserved elements as well as lack of overall conservation [DIETERICH ET AL. 02]. In the current work, first the rat sequences were aligned to orthologous genes in the mouse. Because these species showed a high degree of conservation for the chosen genes, sequences of both were also compared to their human orthologues.

The term orthologous gene is applied to genes that are derived from a common ancestor and separated by speciation [WASSERMAN & SANDELIN 04]. In this study, orthologous genes were retrieved from the Compara database provided by Ensembl. Information stored in Compara can be easily accessed via the Ensembl API and was obtained using a script written by Steffen Grossmann.

Next, an alignment program written by Huang and Miller and extended by C. Dieterich et al. was employed for the detection of CNBs. The program is based on the Waterman-Eggert algorithm, a local alignment algorithm that finds sub-optimal alignments by recalculation of the alignment matrix. This extension of Waterman-Eggert is appropriate in the context of regulatory sequences, since it is both fast and allows assessment of statistical significance [DIETERICH ET AL. 02]. It was run on the fasta files containing the orthologous sequences with the standard parameters (number of alignments: 100, score matrix: Kimura PAM10, gap penalty: 1000, gap extension: 50, probability cut-off: 0.005). Weight matrix scan was done on the significant alignments only.

4.2.4 Screening

From the PSCM to the PSSM

To screen sequences for a certain binding site a model of this binding site is needed. Among the models developed for the description of TFBS two are especially common: the consensus sequence and the position specific count matrix (PSCM).
Both are derived from a multiple alignment of known binding sites, but whereas the consensus sequence can be obtained directly from the PSCM, the PSCM contains more information and cannot be inferred from the consensus sequence.
The PSCM lists for each nucleotide j the number of its occurrences for each position i of the multiple alignment with length L (for the sake of simplicity, it is assumed that all sequences of the alignment have the same length, which is the case for most PSCMs). For an alignment of nucleotide sequences, the PSCM consists of L rows and four columns.
The PSCM is transformed into the position specific frequency matrix (PFM or profile) by dividing each entry by the number of sequences. The entries in the PFM correspond to the probability of finding nucleotide j at position i.
Next, the position specific score matrix (PSSM) is calculated by dividing the profile of the TFBS (signal profile) by the background profile and taking the logarithm. This log-likelihood-ratio defines the score for a given nucleotide at a certain position. The background profile gives the probability of finding nucleotide j at position i in a motif taken from a random sequence. The calculation of the background profile is dependent on the gc-content of the sequence under investigation.

Because zero entries in the signal profile would lead to a logarithm of zero in the PSSM, regularization prior to the calculation of the PSSM is necessary. This is also biological meaningful, as the observation of zero for one nucleotide in a certain position might not be true for all occurrences of the TFBS.

Regularization

In this work, the regularization method introduced by Rahmann and colleagues has been applied [Rahmann et al. 03] as it allows position-dependent regularization and leaves core motifs untouched. Their approach finds the optimal balance between the overall column-wise nucleotide distribution and the nucleotide distribution at each row of the PSCM. Out of a family of regularizing distributions, the best is found by stepwise adjustment of a weight parameter.

Score Scaling

Scaling of the scores was performed as suggested by Rahmann et al. by dividing the scores obtained from the PSSM by 0.05. In addition, scores were rounded to integers. This is of advantage for the computational calculation of score distributions. The natural logarithm has been applied for the calculation of position dependent nucleotide scores.

Score Calculation, False Positives and False Negatives

Sequence screening is done by defining a window of suitable size and shifting it along the sequence base by base. For each subsequence contained in the current window, the score is calculated according to the chosen TFBS model. In case of the PSSM, the individual scores of the nucleotides are summed over the sequence window.

If a sequence window scores higher than a given threshold although it is generated by the background model, it is regarded as false positive (FP). On the other hand, if a motif generated by the signal model scores below the given threshold, it is called a false negative (FN). In case of a PSSM, the amount of expected FN and FP can be derived from a calculation of the score distribution of both the background and the signal model.

The BioMinerva implementation of score distributions is based on the approach outlined by Rahmann and colleagues [RAHMANN ET AL. 03]. The score of all sequences which can be generated by the PSSM is calculated together with its probability under the signal and the background profile. Computation is speeded up by adding up probabilities of scores with identical values in the intermediate steps, thus reducing the amount of possible scores in each step. This allows the calculation of all scores obtainable from the matrix (maximal 4^L, with L as matrix length) together with their probabilities in a sufficiently short time.

With these score distributions at hand, a suitable score threshold for the acceptance of a sequence motif as TFBS can be defined. There are three strategies available for threshold choice:

1) reduction of FPs
2) reduction of FNs
3) balanced number of FPs and FNs

If the reduction of FPs is favored over the detection of true positives, the selectivity of the screen is enhanced at the cost of missing true instances. On the other hand, if one wants to detect as many true positives as possible, the sensitivity of the screen is enhanced at the cost of more FPs. The third strategy balances sensitivity and selectivity. In this study, the first strategy was chosen, because in the context of TFBS screening missing a true instance is less costly than the prediction of a FP, since the verification methods are both time intensive and expensive.

Because many windows are scored for each sequence, the issue of multiple testing arises. To tackle this problem, Rahmann and colleagues distinguish the sequence alpha error probability (α_n) from the window alpha error probability (α). The windows are regarded as statistically independent, although they overlap. This is justified, since only a small number of separated TFBS is expected to occur in the sequence. For a sequence of length n the probability that at least one out of n windows will score higher than the threshold is given by α_n. The probability of a FP for a given number of independent tests follows a Poisson distribution, thus α_n can be defined as:

$$\alpha_n(t) = 1 - (1 - \alpha(t))^n \approx 1 - exp^{(-n\alpha(t))} \tag{4.1}$$

with t denoting the threshold and n the number of windows (notation follows [Rahmann et al. 03]). The selectivity of a given PSCM is then defined as:

$$Q_{sel}(PSCM) = 1 - \alpha_n(t) \tag{4.2}$$

A similar reasoning holds for the false negatives. The occurrence of at least one FN in a sequence with m true instances of the TFBS is given as follows:

$$\beta_m(t) = 1 - (1 - \beta(t))^m \tag{4.3}$$

The sensitivity of a given PSCM follows:

$$Q_{sen}(PSCM) = 1 - \beta_m(t) \tag{4.4}$$

Rahmann et al. suggest the use of n = 500 and m = 1 for multiple test correction. Thus, the threshold is chosen such that in a sequence with 500 bp length one FP is expected for the specified level of significance. The suggested values for m and n were used for the current work with 0.05 as the given level of significance (p-value).

4.2.5 Matrices

Matrix Databases

Position specific count matrices are available from Transfac, which contains a collection of PSCM derived from the literature and Jaspar, a smaller, non-redundant database of PSCMs. The T-Reg database [Roepcke et al. 05] stores both databases locally in a common framework and has been used in this study.

SRF Matrices

In T-Reg, seven matrices describing CArG-boxes were found. Their properties are given in Table 4.2.

Matrix-ID	Consensus sequence	Information content in nats	Source (quality)	Species	Database
M00152	ATGCCCATA TATGGWNNT	16.37	SELEX, 33 selected binding sequences	artificial sequences (SRF from mouse)	TRANSFAC PUBLIC

Matrix-ID	Consensus sequence	Information content in nats	Source (quality)	Species	Database
M00186	GNCCAWATA WGGMN	9.86	21 compiled sequences (Q6)	mixture	TRANSFAC PUBLIC
M00215	DCCWTATAT GGNCWN	10.46	consind generated matrix	mixture	TRANSFAC PUBLIC
M00810	SCCAWATA WGGMN MNNNN	10.27	27 compiled sequences (Q4)	mixture	TRANSFAC
M00922	CCAWATAW GGMNMNG	9.53	29 compiled sequences (Q5)	mixture	TRANSFAC
M01007	CNKNKCCTTA TWTGGNNNN	10.22	54 compiled sequences (Q5)	mixture	TRANSFAC
MA0083	GCCCWTAT AWGG	12.26	SELEX	artificial sequences	JASPAR

Table 4.2: The first column lists the matrix-IDs as stored in T-reg. The consensus sequence displayed in the second column gives for each position the most frequent nucleotide (D = A/G/T, K = G/T, M = A/C, N = any nucleotide, S = G/C, W = A/T). In the third column, the information content of the matrices based on the natural logarithm is shown. The fourth column indicates the data on which the matrix is based, SELEX stands for artificial binding site selection, whereas consind refers to the program ConsIndex that is used for automated matrix generation. For compiled matrices, a quality index is assigned by TRANSFAC with the following meanings: Q4 = known binding sequence, Q5 = binding of uncharacterized extract protein to a binding site, Q6 = no quality assigned.

It is of note that most of the matrices listed in Table 4.2 not only contain the central CArG motif but also a varying number of additional bases. This might reflect the importance of flanking base pairs as mentioned by Mo et al. [MO ET AL.01].

To detect SREs rather than CArG-boxes only, it might be of value to assemble a new matrix which contains only verified SREs. But inspection of the SREs outlined in Table 4.1 demonstrates that distances between CArG-boxes and Ets-sites are in fact too variable for the matrix to be useful in SRE-detection. This variability clearly demonstrates the flexibility of the linker region in TCFs. Therefore, the two binding sites were not combined in a matrix but rather searched for individually.

Ets Matrices

In order to detect SREs the predicted CArG-boxes were screened for Ets-sites 50 bp up- and downstream of the CArG motif, because in most SREs the distance between the CArG-box and the Ets-site does not exceed this value. The parameters chosen for CArG-box detection were also used for Ets motif screening. Vertebrate Ets binding sites are represented in T-Reg with 17 different PSCMs, whose properties are given in the Appendix (Table C).

Matrix Distances

Since many different PSCMs for both the CArG-box and the Ets-site are available, it is desirable to measure the difference between two matrices. If the differences were minor, this would allow to restrict the screen to only one suitable matrix for each binding site. Three different approaches for the measurement of matrix differences have been published recently.

Relative information content [ROEPCKE ET AL. 05]
In this approach, the relative information content (IC_{rel}) is used as similarity measure for two position specific weight matrices.
The IC_{rel} of a matrix A with respect to a matrix B is calculated row-wise as follows:
a) In the region of the overlap: $IC_{rel} = \sum_{i=1}^{O} E_{rel}(A_i, B_i)$
b) Outside of the overlap: $IC_{rel} = \sum_{i=1}^{L} log(4) - E(A_i)$
E_{rel} and E denote the relative entropy and the entropy, whereas O stands for the length of the overlap and L for the length of the non-overlapping region.
Finally, the overlap-dependent values of the IC_{rel} are summed up to obtain the IC_{rel} for matrix A with respect to matrix B.
In order to find the optimal shift between two matrices, IC_{rel} is calculated for all possible shifts (a minimum overlap of both matrices is required). The shift with the largest associated information content is then returned as the optimum. It is of note that this measure is not symmetrical.

Correlation-based similarity measure C [KIEŁBASA ET AL.05, in press]
For the calculation of the Pearson correlation coefficient, a test sequence with random, equi-distributed bases is screened with either of the two PSSMs under investigation. For each of the possible shifts between the two matrices (with minimum overlap of 6 bp) the correlation coefficient of the corresponding scores is calculated. The highest correlation coefficient from all shifts is used as the similarity measure C.

χ^2-based similarity measure D [KIEŁBASA ET AL.05, in press]
For the calculation of the similarity measure D, the nucleotide composition of all possible overlaps (again minimum 6 bp) between two PFMs is compared. To assess similarity of the two base count distributions, the χ^2-test is applied row-wise and the number of significantly different rows is calculated for each shift. The shift resulting in the smallest number of different rows is chosen and its corresponding number of different rows taken as the similarity measure D.

4.3 Results

4.3.1 Promoter Verification

The DBTSS does not indicate a single nucleotide as start site for a given gene, but rather contains a distribution of cDNA start sites. For all genes investigated, the RefSeq TSS was positioned within this range.
In the table below, the average range of cDNA start sites over all genes found in the DBTSS is given:

Species	Mouse	Human
Mean range in bp	197	319
Median range in bp	105	176

In a study covering 276 human genes, an average range of start sites of 62 bp was obtained [SUZUKI ET AL.01]. The authors also mention that genes can be divided in two classes according to their range of start sites: those with tightly clustering start sites and those with highly variable start sites. The large value for human genes found in the present study is due to mainly two genes with exceptional wide ranges (guanine nucleotide binding protein and jak1). To obtain a less biased average range, the median range was also calculated.

The table below lists start sites stored in the EPD in comparison with those from Ensembl. The TSS is given in absolute chromosomal coordinates. Only a small subset of the 25 genes under investigation could be found in the EPD.

Gene (Chromosome)	Species	Ensembl-ID	TSS given by Ensembl	TSS given by EPD
Antioxydant enzyme (Chr19)	human	ENSG00000167815	12773694	12773683
HNOP56 (Chr20)	human	ENSG00000101361	2581254	2581278
ODC-1 (Chr2)	human	ENSG00000115758	10539051	10539090
ODC-1 (Chr12)	mouse	ENSMUSG00000011179	17672074	17672074
P53 (Chr17)	human	ENSG00000141510	7531642	7531677
Ppp1 (Chr11)	human	ENSG00000172531	66925952	66925903

For those genes where comparison of both databases was possible, the start sites correspond well. In average, the TSS given by the EPD differs less than 30 bp from the TSS stored in Ensembl.

4.3.2 Matrix Selection

Matrix distances were measured with the three methods described in section 4.2.5 and are given in Figures 4.2-4.5. The tree structures in Figure 4.2 and 4.4 were obtained from http://wmcompare.gene-groups.net.

SRF

The two matrices derived from SELEX experiments (M00152, MA0083) are indicated as closest to each other by all three methods. In the tree, they are separated from the other matrices (see Figure 4.2) and form their own cluster. In agreement with these findings, the distance between these matrices is encoded by light colors in Figure 4.3 (see upper left and lower right corner).

The matrices based on known binding sites form a cluster of relatively high similarity. From this cluster, M00810 was chosen, because it has a high quality, a high information content and is derived from a comparatively high number of validated binding sites. Its composition cannot be published in this work, since M00810 is stored in the commercial part of TRANSFAC.

The two SELEX matrices with their high information content were not considered for screening as it is unclear whether results obtained from artificial binding sites are biologically meaningful.

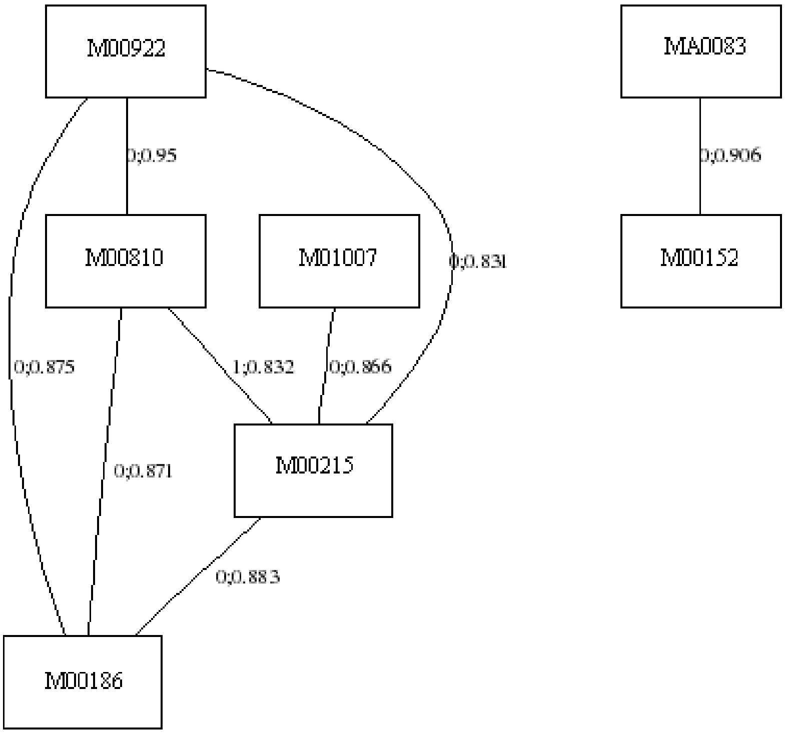

Figure 4.2: Here, the tree structure resulting from similarity measures D and C is shown for the SRF matrices. Missing links between the matrices indicate C and D values below the threshold ($C \geqq 0.8$, $D < 2$). It is of note that the regularization of PSCMs differs from the approach used for matrix screening. The tree was obtained with the help of software written by Kiełbasa and colleagues.

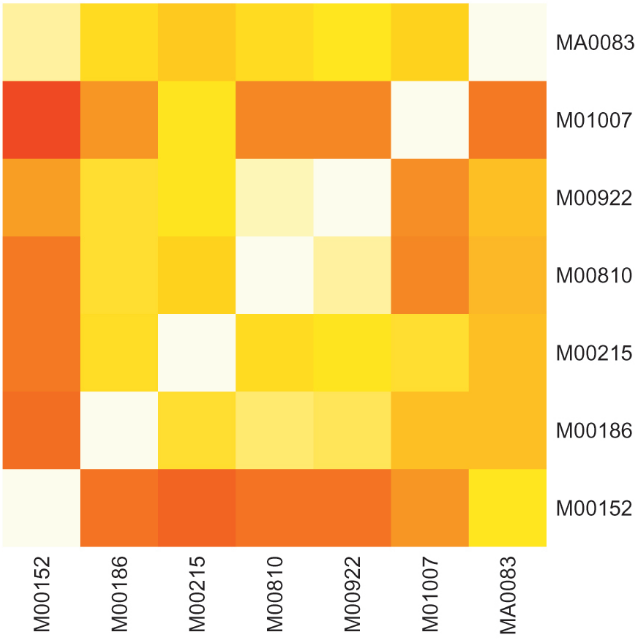

Figure 4.3: IC_{rel} dissimilarity matrix of SRF matrices is given with colors ranging between white (high IC_{rel}) and red (low IC_{rel}).

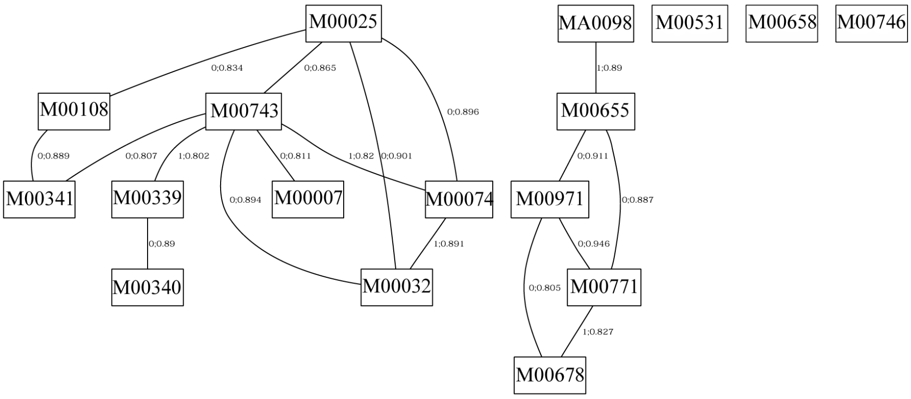

Figure 4.4: Here, the tree structure resulting from similarity measures D and C is shown for the Ets matrices. Missing links between the matrices indicate C and D values below the threshold (C $\geqq$ 0.8, D < 2). It is of note that the regularization of PSCMs differs from the approach used for matrix screening. The tree was obtained with the help of software written by Kiełbasa and colleagues.

Ets

The Ets matrices are separated into two clusters and three single matrices (see Figure 4.4). This result agrees well with the dissimilarity matrix (see Figure 4.5). If for example M00971 is investigated in the dissimilarity matrix, its closest neighbors (M00678, M00771, M00655 and MA0098) are found to be connected with it in the tree.

For Ets-screening, restriction to one matrix is not necessary, since the screened sequences are very short. The larger of the two clusters is more interesting, because it contains two matrices for the known ternary complex factor Elk-1 (M00007, M00025). Members of this cluster show different degrees of similarity, so selecting only one matrix might reduce detection power. Therefore, all members of the Elk-1 cluster were chosen for screening.

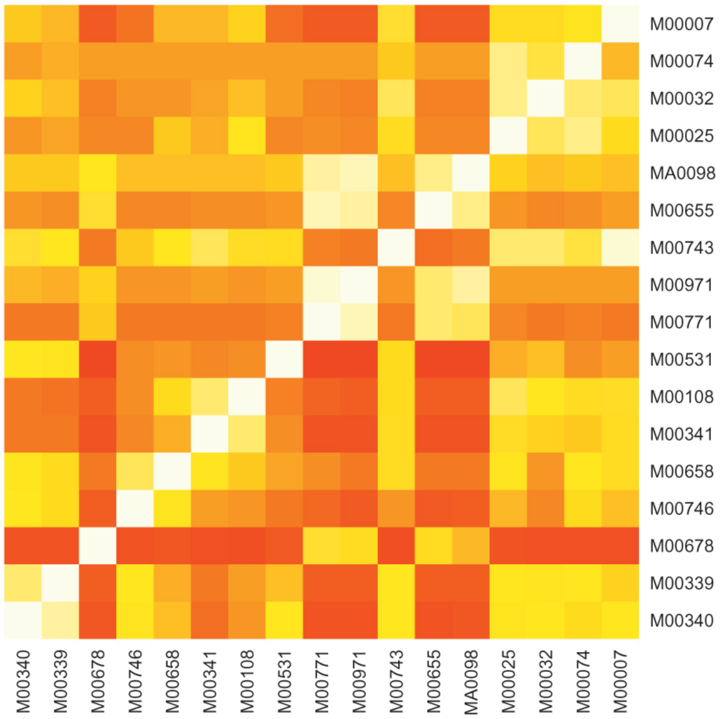

Figure 4.5: IC_{rel} dissimilarity matrix of Ets matrices is given with colors ranging between white (high IC_{rel}) and red (low IC_{rel}).

4.3.3 SRF Target Screening

Conservation

The table below gives the degree of conservation expressed as the percentage of sequence length left after the alignment. This percentage is averaged over all investigated genes.

Alignment	Mouse vs Rat	Mouse vs Human	Rat vs Human
upstream	62.7 (mouse)	16.9 (mouse)	14 (rat)
	64.2 (rat)	15.8 (human)	13 (human)
downstream	68.1 (mouse)	29.9 (mouse)	28.1 (rat)
	68.9 (rat)	30.4 (human)	28.6 (human)

It is of note that downstream regions are better conserved than upstream regions. This might be due to the first intron, which is included in the upstream region and which might be less conserved than the downstream region.
In addition, sequence length is not substantially reduced by mouse/rat alignments. This demonstrates that mouse and rat are too closely related for effective phylogenetic footprinting.

Promising SRF Target Candidates

Screening for SRF binding sites revealed a number of promising target genes. As the conservation between mouse and rat is high, only those genes are chosen as likely candidates that contained at least one CArG-box conserved in rodent and human sequences. PDGF-A as the only exception will be discussed in section 4.4. The winning hits are described in detail in Table 4.3 below.

Gene, number of hit (alignment)	Position of hit (strand), distance to TSS in bp	Length of all CNBs screened in bp (FPs)	Hit-CNB: percent-id, gc-content	Hit: score (cut-off), sensitivity	Sequence of CArG-Box
Alpha actinin, 1. hit (**mouse** vs human)	Chr7: 17900584-17900601 (-), +30739	12219 (up) 1215 (down) (27)	72, 0.49	239 (129), 0.91	CCTTATATGG
Alpha actinin, 2. hit (**mouse** vs rat)	Chr7: 17904603-17904620 (+), +26720	44107 (up) 4503 (down) (97)	47, 0.47	154 (129), 0.91	CCAAAAATGG

Gene, number of hit (alignment)	Position of hit (strand), distance to TSS in bp	Length of all CNBs screened in bp (FPs)	Hit-CNB: percent-id, gc-content	Hit: score (cut-off), sensitivity	Sequence of CArG-Box
Alpha actinin, 3. hit (**mouse** vs human)	Chr7: 17914645-17914662 (-), +16678	12219 (up) 1215 (down) (27)	63, 0.51	131 (129), 0.91	CCTAATAAGG
Alpha actinin, 1. hit (**rat** vs human)	Chr1: 84131115-84131132 (-), +20655	11205 (up) 1628 (down) (26)	70, 0.49	153 (129), 0.91	CCTAATAAGG
Alpha actinin, 2. hit (**rat** vs mouse)	Chr1: 84115517-84115534 (-), +36253	44569 (up) 4623 (down) (98)	47, 0.48	206 (129), 0.91	CCAAAAAAGG
Alpha actinin, 3. hit (**rat** vs human)	Chr1: 84117050-84117067 (-), +34720	11205 (up) 1628 (down) (26)	71, 0.51	265 (129), 0.91	CCTTATATGG
Alpha actinin, 1. hit (**human** vs mouse)	Chr19: 43848730-43848747 (+), +18563	12701 (up) 1238 (down) (28)	63, 0.53	162 (129), 0.91	CCTAATAAGG
Alpha actinin, 2. hit (**human** vs mouse)	Chr19: 43842314-43842331 (-), +12147	12701 (up) 1238 (down) (28)	74, 0.44	208 (127), 0.92	CCAAATAAGG
Alpha actinin, 3. hit (**human** vs mouse)	Chr19: 43866218-43866235 (+), +36051	12701 (up) 1238 (down) (28)	72, 0.52	240 (129), 0.91	CCTTATATGG
Cpg21, 1. hit (**mouse** vs human)	Chr19: 52886096-52886113 (-), -262	3513 (up) 1043 (down) (9)	58, 0.7	274 (115), 0.96	CCATATTTGG

Gene, number of hit (alignment)	Position of hit (strand), distance to TSS in bp	Length of all CNBs screened in bp (FPs)	Hit-CNB: percent-id, gc-content	Hit: score (cut-off), sensitivity	Sequence of CArG-Box
Cpg21, 2. hit (**mouse** vs human)	Chr19: 52886102-52886119 (+), -256	3513 (up) 1043 (down) (9)	58, 0.7	253 (115), 0.96	CCTTATATGG
Cpg21, 1. hit (**rat** vs human)	Chr1: 259965196-259965213 (+), -83	3559 (up) 678 (down) (7)	63, 0.65	228 (121), 0.94	CCATATTTGG
Cpg21, 2. hit (**rat** vs human)	Chr1: 259965200-259965217 (-), -79	3559 (up) 678 (down) (7)	63, 0.65	299 (121), 0.94	CCTTATATGG
Cpg21, 1. hit (**human** vs mouse)	Chr10: 112247568-112247585 (-), -98	3569 (up) 981 (down) (9)	58, 0.74	243 (107), 0.98	CCATATTTGG
Cpg21, 2. hit (**human** vs mouse)	Chr10: 112247574-112247591 (+), (-92)	3569 (up) 981 (down) (9)	58, 0.74	283 (107), 0.98	CCTTATATGG
JunB, 1. hit (upstream) (**mouse** vs human)	Chr8: 84255704-84255721 (+), -1729	2644 (up) 1201 (down) (8)	63, 0.57	211 (127), 0.92	CCATATTAGG
JunB, 2. hit (downstream) (**mouse** vs rat)	Chr8: 84252169-84252186 (+), +1806	4919 (up) 3628 (down) (17)	56, 0.52	245 (129), 0.91	CCATATATGG
JunB, 1. hit (upstream) (**rat** vs mouse)	Chr19: 24834889-24834906 (-), -1743	4911 (up) 3904 (down) (18)	76, 0.56	192 (128), 0.92	CCATATTAGG

Gene, number of hit (alignment)	Position of hit (strand), distance to TSS in bp	Length of all CNBs screened in bp (FPs)	Hit-CNB: percent-id, gc-content	Hit: score (cut-off), sensitivity	Sequence of CArG-Box
JunB, 2. hit (downstream) (**rat** vs human)	Chr19: 24838425-24838442 (+), +1793	2685 (up) 1051 (down) (8)	81, 0.71	290 (112), 0.95	CCATATATGG
JunB, 1. hit (upstream) (**human** vs mouse)	Chr19: 12761755-12761772 (-), -1555	2774 (up) 1241 (down) (8)	63, 0.7	220 (115), 0.96	CCATATTAGG
MKP-3, 1. hit (**mouse** vs human)	Chr10: 99042450-99042467 (+), -1915	3582 (up) 2362 (down) (12)	64, 0.58	197 (128), 0.92	CCTTTTTTGG
MKP-3, 2. hit (**mouse** vs human)	Chr10: 99042534-99042551 (+), -1831	3582 (up) 2362 (down) (12)	64, 0.58	159 (128), 0.92	CCAATTTTGG
MKP-3, 1. hit (**rat** vs human)	Chr7: 36911714-36911731 (+), -2254	3470 (up) 2517 (down) (12)	66, 0.58	172 (128), 0.92	CCAATTTTGG
MKP-3, 2. hit (**rat** vs human)	Chr7: 36911630-36911647 (+), -2338	3470 (up) 2517 (down) (12)	66, 0.58	196 (128), 0.92	CCTTTTTTGG
MKP-3, 1. hit (**human** vs mouse)	Chr12: 88250796-88250813 (-), -2032	3793 (up) 2371 (down) (12)	64, 0.61	154 (125), 0.93	CCATTTTTGG
MKP-3, 2. hit (**human** vs mouse)	Chr12: 88250712-88250729 (-), -1948	3793 (up) 2371 (down)	64, 0.61	173 (125), 0.93	CCAAAATTGG

Gene, number of hit (alignment)	Position of hit (strand), distance to TSS in bp	Length of all CNBs screened in bp (FPs)	Hit-CNB: percent-id, gc-content	Hit: score (cut-off), sensitivity	Sequence of CArG-Box
PDGF-A, 1. hit (**mouse** vs rat)	Chr5: 136399972-136399989 (+), -1185	4876 (up) 4272 (down) (18)	46, 0.63	231 (122), 0.94	CCTTTTATGG
PDGF-A, 1. hit (**rat** vs mouse)	Chr12: 16182227-16182244 (-), -3189	4951 (up) 4441 (down) (19)	46, 0.43	224 (129), 0.92	CCTTTTATGG

Table 4.3: This table displays detailed information on the CArG-boxes of candidat genes. For each gene, the hits found in each species are listed. Double horizontal lines separate different species.
The first column indicates the number of the hit in the putative target gene for the given species. The alignment is shown in parentheses, with the screened species printed in bold. As far as possible, the hits obtained by alignments of rodent/human sequences are listed.
The position of the hits is given in the second column in absolute chromosomal coordinates. The distance relative to the TSS is also shown. In this context, a minus (-) stands for an upstream position of the CArG-box relative to the TSS and a plus (+) for a position downstream of the TSS. It is of note that start and end position refer not to the CArG-box but to the motif encoded in the matrix used for screening (M00810).
In the third column, the total length of the sequence screened for the given gene is shown, separated in the length of the up- and downstream region. In parentheses, the number of expected false positives with the chosen probability of $\alpha_n = 0.05$ is also listed.
The first number in the fourth column describes the degree of conservation of the CNB where the hit was found (identical bases in percent), whereas the second number represents its gc-content. This gc-content was used to calculate the cut-off score, which is given in the fifth column in parentheses. Because α_n was fixed at 0.05, the selectivity is constant (0.95).

Table 4.3. demonstrates that increasing gc-content improves sensitivity. In a region with high gc-content, it is easier to recognize the CArG-box with its gc-content of 0.4. Consequently, the threshold can be decreased leading to a reduced loss of true instances. The same holds for low gc-contents (below 0.4), which were not observed in the CNBs investigated.

In Figure 4.6. the signal and background score distributions are shown, which were used to calculate the cut-off score dependent from the given gc-content.

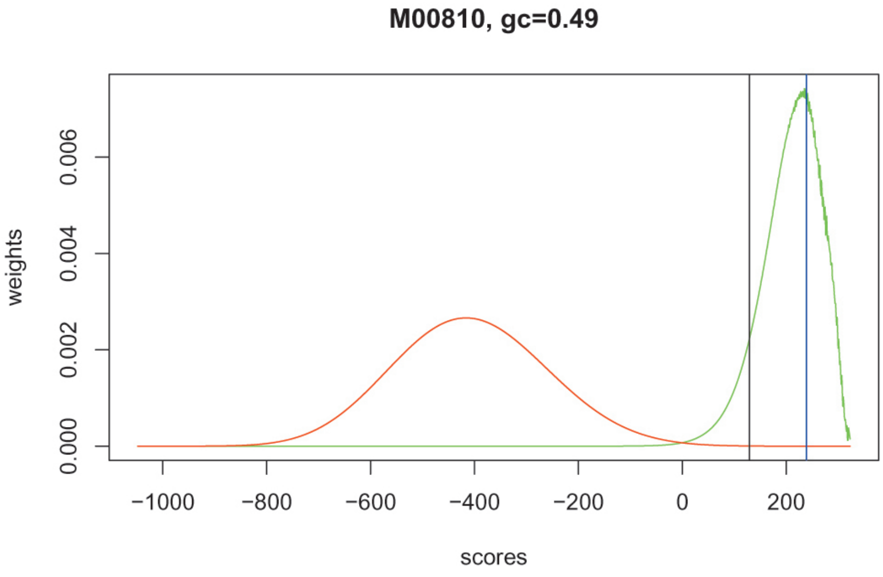

Figure 4.6: The signal score distribution is colored green, whereas the background distribution is shown in red. The score of the hit (found in murine alpha actinin) is marked by a blue vertical line, the cut-off score by a black one.

CArG-boxes were also found in the genes helicase p68, ks-1, tdag51, ODC-1 and jak-1, but only in rodents. The absence of these CArG-boxes in the human genes might be either due to differences in the biology of humans and rodents or due to the fact that they are false positives. Because of the latter reason, these CArG-boxes are not discussed in detail.

In Figures 4.7-4.11, each promising target gene is shown in its genomic context. In these Figures, the CNBs are displayed as green bars. The gene as predicted by Ensembl

is shown together with its exons (orange boxes), introns (black lines) and different transcripts (if more than one exist). Surrounding genes are also shown together with their transcripts. Genes and transcripts are labeled with their Ensembl-ID. The arrow indicates whether the gene is positioned on the plus or minus strand. Below the green CNBs, the position of the hits is marked by small black boxes. The scales above and below the genes give the absolute chromosomal coordinates.
Interestingly, in the five predicted target genes the conserved regions are scattered over the whole sequence range and not centered near the TSS.

Of the three known SRF targets (junB, cpg21, PDGF-A), two were found to have CArG-boxes in all three species. In PDGF-A, the CArG-box is detected in rat and mouse, but not in human. Since the promoter region of the human PDGD-A was published [Lin et al. 92], an attempt was made to map its CarG-box onto the human genome with the blat program provided by the UCSC genome browser. The reason for the failure of this mapping is discussed in section 4.4.

4.3.4 Detection of SREs

In order to test whether predicted CArG-boxes belong to SREs, short regions flanking the CArG-boxes were screened for Ets motifs. The result of this screen is given in Table 4.4.

Gene	Species, number of predicted CArG-box	Ets matrices (score)	Position of hit (strand), distance to CArG-box in bp	Sequence
Alpha actinin	Human, 2. CarG-box	M00339 (8.3)	43842368-43842382 (-), +51	GGAGGAAGTTTTGCA
JunB	Mouse, upstream CarG-box	1) M00341 (7.95) 2) M00180 (9,7)	1) 84255717-84255728 (+), +13 2) 84255717-84255726 (+), +13	1)ACAGGAAGAGGT 2)ACAGGAAGAG
JunB	Rat, upstream CarG-box	1) M00341 (7.8) 2) M00108 (9.7)	1) 24834882-24834893 (-), -13 2) 24834884-24834893 (-), -13	1)ACAGGAAGAGGT 2)ACAGGAAGAG

Gene	Species, number of predicted CArG-box	Ets matrices (score)	Position of hit (strand), distance to CArG-box in bp	Sequence
JunB	Human, upstream CarG-box	1) M00341 (9.15) 2) M00108 (10.7)	1) 12761748-12761759 (-), -13 2) 12761750-12761759 (-), -13	1)ACAGGAAGAGGT 2)ACAGGAAGAG

Table 4.4: Ets-sites within a 50 bp distance to the given CArG-boxes are listed. In the third column, the matrices that detected an Ets-site are shown together with the scores of their hits.
The position of the Ets-sites in absolute chromosomal coordinates is displayed in the fourth column. If the Ets-site is positioned upstream relative to the CArG-box, the relative distance is given together with a minus (-), otherwise with a plus (+).

Ets screening failed to detect new Ets motifs near the CArG-boxes of the putative SRF target genes. The known Ets-site of junB was verified, whereas the hit found in human alpha actinin is not conserved and therefore doubtful.
It is of note that in junB CArG-box and Ets motif have the same orientation, even if it is opposed to the orientation of the gene. (The orientation of the CArG-box could be determined although it is nearly palindromic, because the matrix used for screening not only covers the CArG-box but also surrounding nucleotides. Therefore, slightly different scores were obtained for CArG-boxes on different strands.)
It can be derived from Table 4.3 and 4.4 that the order of CArG-box and Ets-site is fixed in junB, with the Ets-site upstream of the CArG-box with regard to the TSS. This order is also observed in mcl-1, c-fos and two of the three SREs in egr-1 (Table 4.1).
Another striking fact is that not the Elk-1 matrices but the matrices for the related Ets family members GA-binding protein and nuclear respiratory factor 2 detected the junB Ets-site.

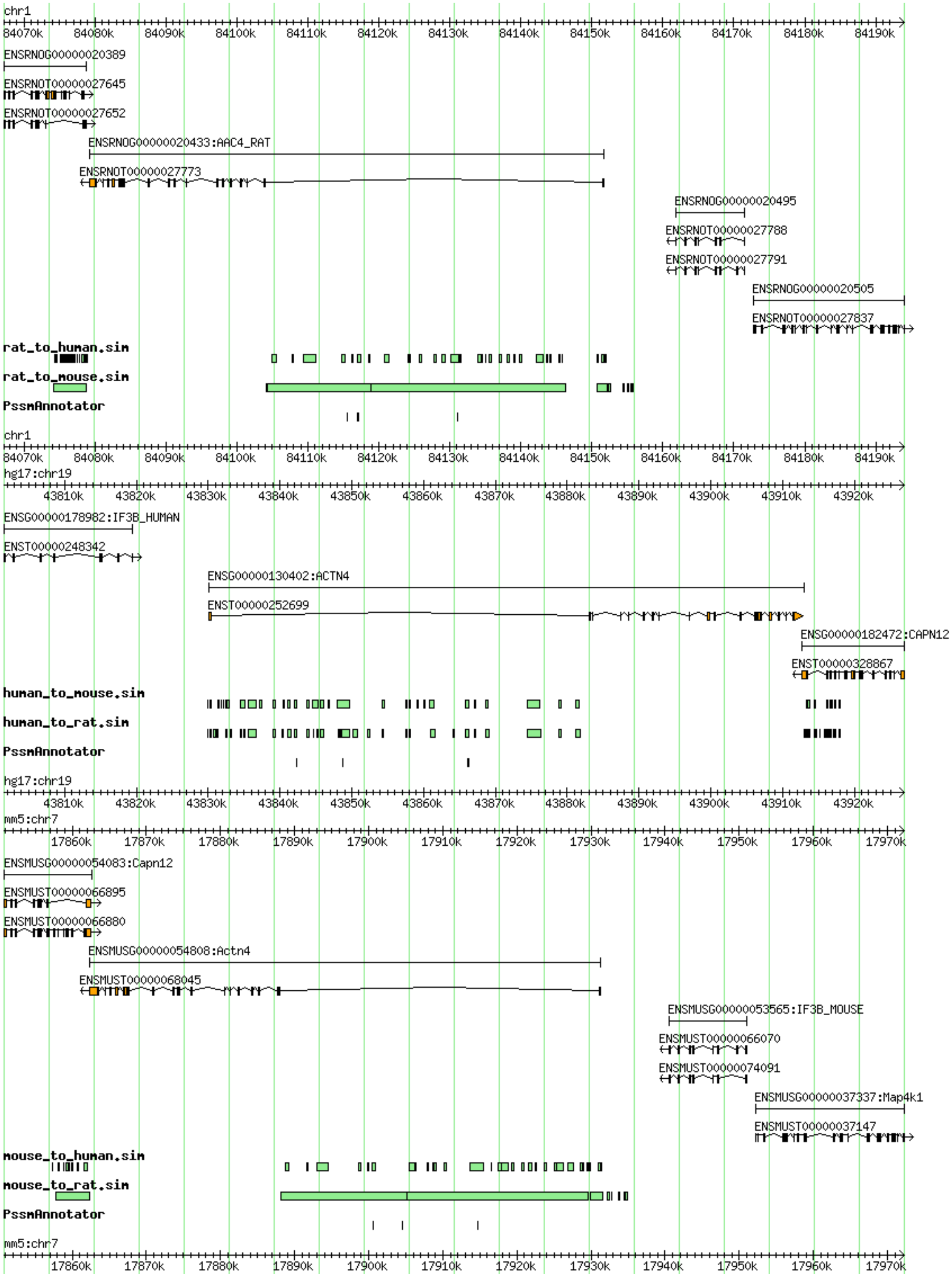

Figure 4.7: The genomic context of alpha actinin is shown for all three species (from top to bottom: rat, human, mouse). Green bars represent CNBs, orange bars exons and the black boxes below the CNBs the hits.

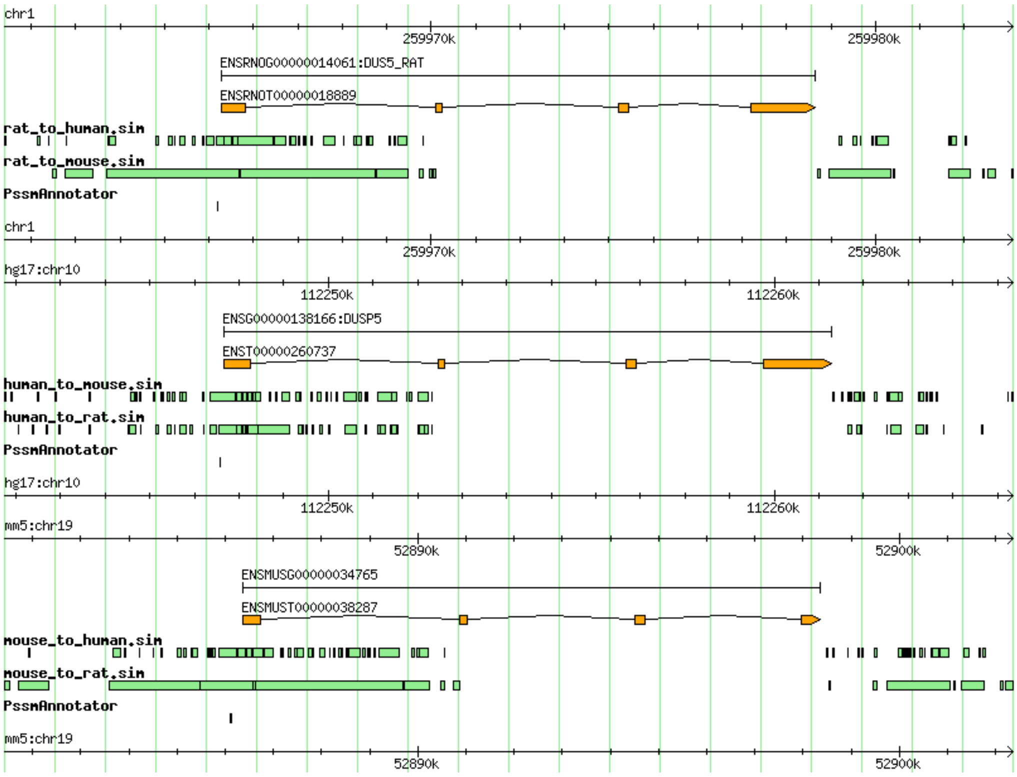

Figure 4.8: The genomic context of cpg21 is shown for all three species (from top to bottom: rat, human, mouse). Green bars represent CNBs, orange bars exons and the black boxes below the CNBs the hits. The two hits are positioned so close to each other that only one hit can be seen on the picture.

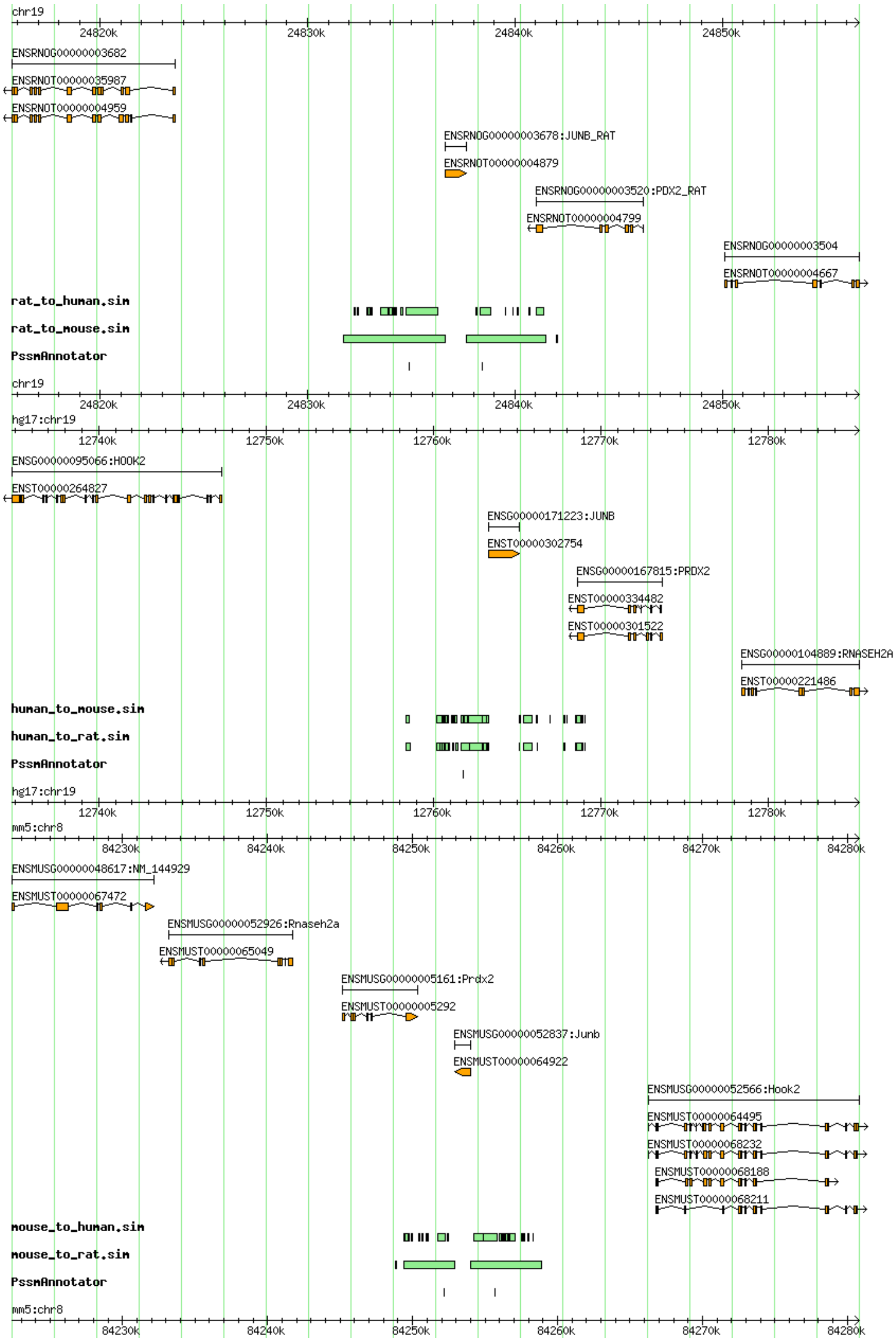

Figure 4.9: The genomic context of junB is shown for all three species (from top to bottom: rat, human, mouse). Green bars represent CNBs, orange bars exons and the black boxes below the CNBs the hits. It is of note that the neighbor gene downstream from junB is antioxidant enzyme AOE372, which is also a member of cluster B.

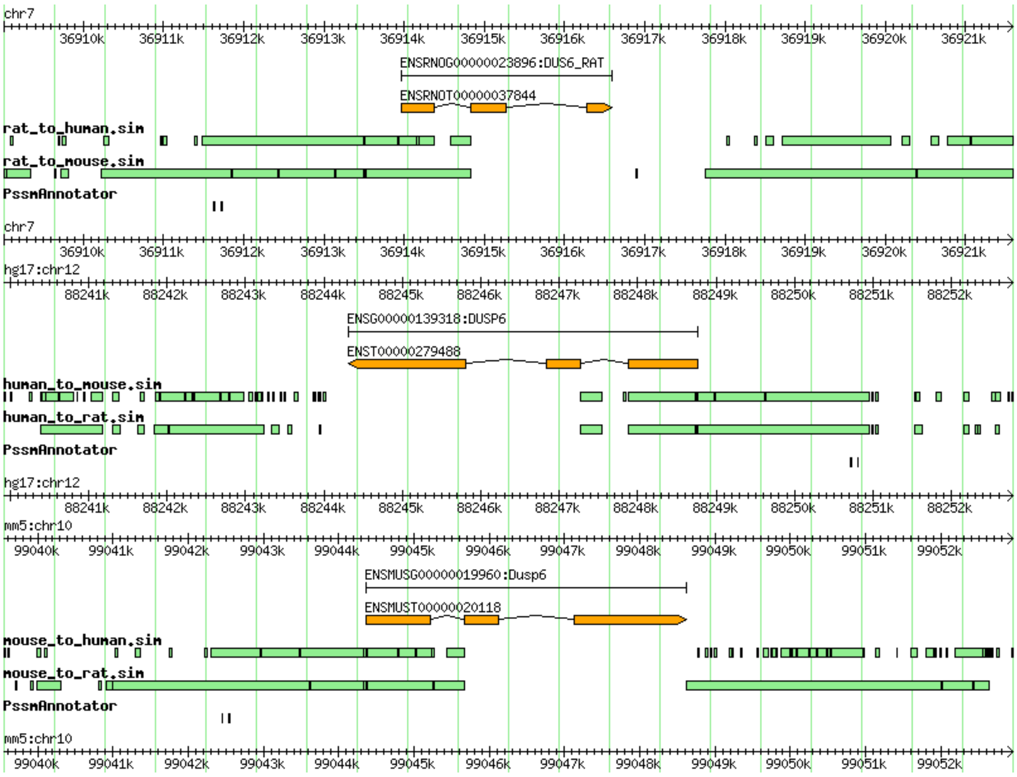
chr7
36910k 36911k 36912k 36913k 36914k 36915k 36916k 36917k 36918k 36919k 36920k 36921k
ENSRNOG00000023896:DUS6_RAT
ENSRNOT00000037844
rat_to_human.sim
rat_to_mouse.sim
PssmAnnotator
chr7
36910k 36911k 36912k 36913k 36914k 36915k 36916k 36917k 36918k 36919k 36920k 36921k
hg17:chr12
88241k 88242k 88243k 88244k 88245k 88246k 88247k 88248k 88249k 88250k 88251k 88252k
ENSG00000139318:DUSP6
ENST00000279488
human_to_mouse.sim
human_to_rat.sim
PssmAnnotator
hg17:chr12
88241k 88242k 88243k 88244k 88245k 88246k 88247k 88248k 88249k 88250k 88251k 88252k
mm5:chr10
99040k 99041k 99042k 99043k 99044k 99045k 99046k 99047k 99048k 99049k 99050k 99051k 99052k
ENSMUSG00000019960:Dusp6
ENSMUST00000020118
mouse_to_human.sim
mouse_to_rat.sim
PssmAnnotator
mm5:chr10
99040k 99041k 99042k 99043k 99044k 99045k 99046k 99047k 99048k 99049k 99050k 99051k 99052k

Figure 4.10: The genomic context of MKP-3 is shown for all three species (from top to bottom: rat, human, mouse). Green bars represent CNBs, orange bars exons and the black boxes below the CNBs the hits.

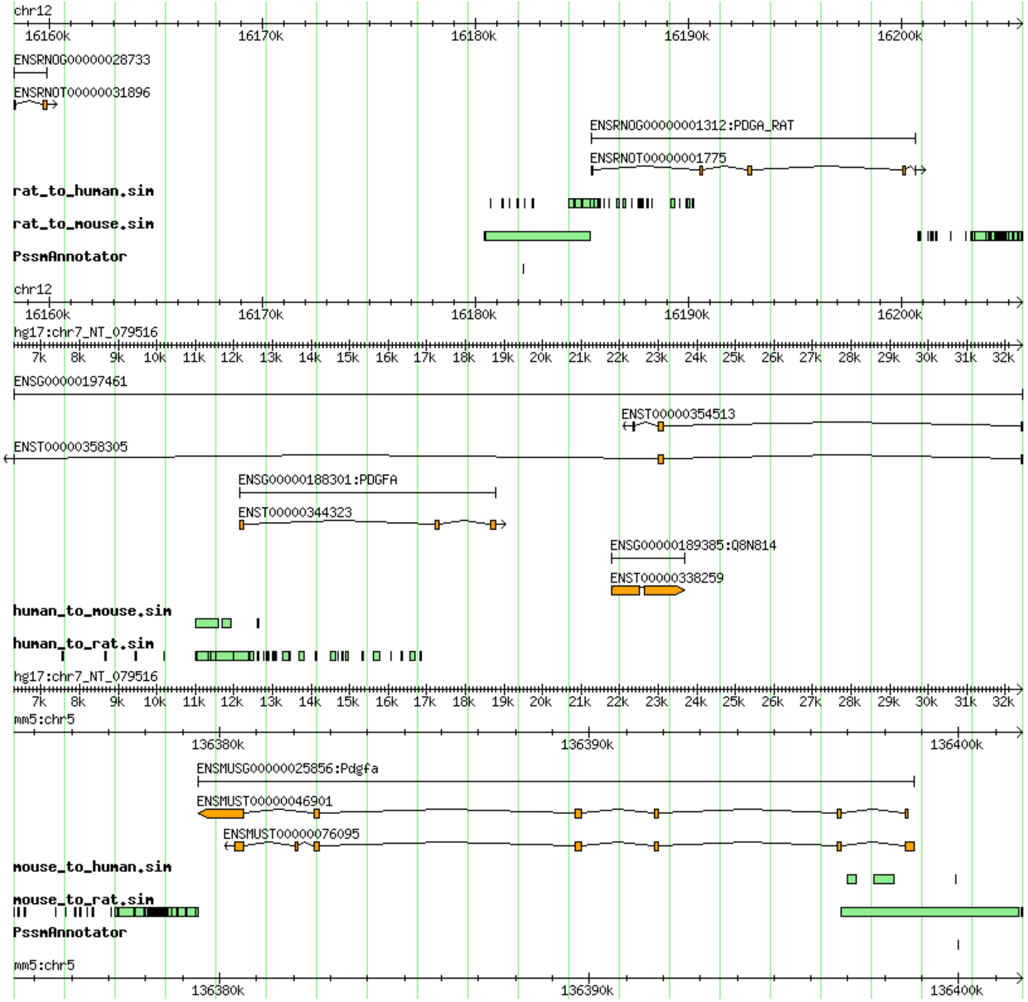

Figure 4.11: The genomic context of PDGF-A is shown for all three species (from top to bottom: rat, human, mouse). Green bars represent CNBs, orange bars exons and the black boxes below the CNBs the hits. It is of note that human PDGF-A is overlapped by an unknown long gene (Ensembl-ID ENSG00000197461).

4.4 Discussion

Problems of the Annotation Procedure

Large gaps between the start of translation and the start of transcription as given by the EPD can occur [DIETERICH ET AL. 02]. These regions could contain TFBS, but were not investigated in the present study. Therefore, TFBS positioned in this region would be missed.

Another issue is the exact position of the TSS. The TFBS screen presented in this work is based on Ensembl TSS definitions and does not take into account alternative TSS given by other sources. As has been noted by Suzuki et al. [SUZUKI ET AL. 01], there might be no fixed TSS at all. For future screening, a sequence region derived from promoter databases rather than a single base pair should be used as TSS.

Disturbingly, mapping of the PDGF-A CArG-box onto the human genome failed. This might be explained by the fact that human PDGF-A is positioned on a contig (7_NT_079516) that is currently not included in chromosome 7 and therefore not detectable with the UCSC genome browser.
Although the position of PDGF-A in the human genome is unknown, the sequence itself is known. Therefore, it is problematic that the CArG-box described in [LIN ET AL. 92] could not be found. This could be due to either the prediction process (false negative) or due to differences in the sequence described by Lin et al. and published in Ensembl.

An additional problem is the filtering step, which was introduced to reduce the amount of detected hits. From all hits found, only those containing the classical CArG-box ($CC(A/T)_6GG$) are listed in section 4.3.3. Therefore, the screening procedure consisted of two steps: First, hits were detected using a matrix model. From the number of significant hits thus obtained, some were chosen according to biological knowledge, in the hope to reduce the number of false positives. This knowledge consists of site-directed mutation studies [MIANO 03] and the crystal structure of the complex [MO ET AL. 01], which show that only small deviations from the classical motif are possible.
Among the discarded hits, a deviating CArG-box found downstream of the human junB gene is especially interesting. It is conserved in mouse and rat with CArG-boxes matching the motif perfectly. This CArG-box seems to represent a functioning, deviating SRF-binding site. Due to the filtering step, other deviating but biological active CArG-boxes might be missed.

Another interesting matter is the orientation of CArG-box and Ets motif. Prediction of the orientation of the (nearly) palindromic CArG-box is only possible if surrounding

nucleotides are taken into account, which introduce a slight difference between scores on the plus and minus strand.
If the order of both motifs were fixed, the orientation of the SRE could be determined by the orientation of the Ets motif. However, at least in egr-1 one SRE is known that deviates from the order observed in the other SREs described in this chapter. Given these facts it is currently not possible to predict the relative orientation of CArG-box and Ets-motif with certainty.

Candidate SRF Targets

Analysis of cluster B revealed a number of putative SRF targets. Clearly, this cluster is enriched with SRF target genes, since three out of 25 genes are known SRF targets.
These three genes (PDGF-A, junB and cpg21) provide a positive control for the method. If they had not been found, this would have pointed to an unfavorable parameter choice, allowing too many false negatives.

Alpha actinin

Three CArG-boxes were found in conserved regions in the first intron of alpha actinin for all three species investigated. Nonetheless, the biological relevance of these CArG-boxes is doubtful. Because alpha actinin is long, the number of expected false positives in the sequence is high. To get an idea about the importance of the CArG-boxes in alpha actinin, the investigation of their conservation in more distantly related species would be helpful.

Cpg21

As described above, cpg21 was demonstrated to be an SRF target by Tullai et al. and therefore is one of the positive controls. Unfortunately, these authors do not comment on the CArG-boxes in cpg21 in detail. This target is exceptional for having two CArG-boxes placed next to each other without an intermediate nucleotide. The CArG-boxes are well conserved in mouse, rat and human and are situated upstream near the TSS. A serum response element might be missing, as no Ets motif could be found.

JunB

JunB, a well known SRF target, is another positive control. The upstream CArG-box as well as the Ets motif described in the literature could be detected in all three species. The downstream CArG-box was found in both rodents and a deviated downstream CArG-box in human. Ets motif screening detected the known Ets sites in all three species (sometimes on the opposite strand).
Interestingly, antioxydant enzyme, another member of cluster B, is situated next to junB, so the downstream CArG-box of junB could as well be assigned to the downstream region of antioxydant enzyme. Functionality of the downstream CArG-box

for murine junB was demonstrated by Perez-Albuerne and colleagues. It would be interesting to know whether expression of antioxydant enzyme is affected by a knockout of this CArG-box.

MKP-3

MKP-3 is the best candidate predicted by this survey. Each of the three species investigated contains two CArG-boxes about 2000 bp away from the TSS. MKP-3 was missed by Tullai et al. because these authors restricted their screen to a window size of 1000 bp upstream of the promoter. Interestingly, two other members of the DUSP-family, namely cpg21 and DUSP2, are already known or predicted SRF targets. It is of note that MKP-3 was not found in the knockout experiment performed by Philippar and colleagues. Its expression level might have been too low to detect differential expression in the chosen cell line.

PDGF-A

As has been mentioned before, hits could be found only in rat and mouse, although human PDGF-A is a known SRF target. The two rodent hits are situated 1000-2000 bp upstream of the TSS. In human, the PDGF-A gene is overlapping with another gene with the Ensembl ID ENSG00000197461 (see Figure 4.11). No information concerning the function of this gene was available. It is no longer present in the newest version of Ensembl.

5 Conclusion

In the course of this work, data derived from a microarray experiment was clustered using a number of different methods. In addition, one cluster was screened for the presence of SRF targets. As a result, four clusters were obtained and MKP-3 was predicted as a new SRF-candidate.

The enrichment of cluster B with known SRF targets points to the biological relevance of the clusters found and supports the assumption that cluster analysis can detect co-regulated genes.
It is of great interest that conserved CArG-boxes in MKP-3 were detected, since this gene is part of the negative feedback loop regulating MAPK. If MKP-3 were indeed up-regulated by SRF, this would provide a link between Ras signaling and MAPK inactivation by DUSPs.

Due to time constraints, a number of interesting questions could not be answered.
First, the question arises whether members of the other clusters share binding sites for another transcription factor. Then, it would be useful to test either the given gene set or a number of randomly chosen genes for the presence of CArG-boxes in order to confirm over-representation of CArG-boxes in cluster B.
It would be also of interest whether human and rodents really represent the optimal evolutionary distance for the investigation of SRF targets or whether more distantly related species would be more suitable.
Another point is the optimization of parameters for clustering (SOM) and weight matrix scan (score threshold). These parameters were chosen according to recommendations in the literature, but it is not clear whether they are the best choice for the data set.

Outlook

In the next step, knockout of SRF in the IR-4 cells for example with siRNA could indicate whether MKP-3 indeed is a SRF target. If up-regulation of MKP-3 in SRF deficient cells fails, this would give the first experimental evidence for the involvement of SRF in MKP-3 regulation.

Another way to confirm putative SRF targets on a large scale is the ChIP-on-Chip analysis, which combines chromatin immunoprecipitation with microarrray technology. An immunoprecipitation-based approach could also test whether Ets proteins like Elk-1 can bind to the MKP-3 promoter, which would point to the formation of a ternary complex.

Recently, the results of a new mapping of human transcription start sites were published [KIM ET AL. 05]. The authors obtained 12,150 binding sites of the transcription factor IID by performing ChIP-on-chip analysis of the whole genome. These were matched to the 5'-ends of known transcripts stored in RefSeq, GenBank and Ensembl. Only 83 % of them were found within 500 bp of the TSS given by these databases. This new data set could not be taken into consideration in the current work, but can be used to improve gene models and to enhance accuracy of prediction in future transcription factor target screenings.

In his model of the MAPK cascade, Nils Blüthgen included the negative feedback loop exerted by the MKPs (Blüthgen, in preparation). The data derived from this model fit well the observed behavior of Ras, P-ERK and MKP-3 as monitored by Western blots and the microarray experiment. However, it could be seen that nuclear MKPs (MKP-1 in cluster A, cpg21 in cluster B) and cytosolic MKPs (MKP-3, MKP-4, both in cluster B) behave differently. Therefore, an extension of this model with respect to the different MKPs would be interesting.

In experiments comparing gene expression in immortalized with those in Ras-transformed rat fibroblasts, about 244 genes were found to be differentially expressed [ZUBER ET AL. 00].
However, the microarray experiment investigated in this study detected only 82 significantly differentially expressed genes. In addition, only a few of these genes exceed twofold differential expression. Therefore, further microarray experiments might not only verify the present results but also allow the analysis of additional genes absent in the current data set.

A Accession Numbers

Table A lists the 82 significantly differentially expressed genes together with their accession numbers (GenBank-IDs).

Gene name	Accession number
Alpha actinin	U19893
antioxidant enzyme AOE372	NM_017169
Arp3	AF307852
aryl hydrocarbon receptor (AHR)	NM_013149
balbc aldose reductase-related protein	AF182168
BCSC-1	AF002672
CAMK-related peptide	AF045469
cAMP-dependent protein kinase type II (prkar2b)	M12492
cca1	AB000215
c-myc	M23418
Cox2	S67722
CSF-1	M84361
cytocentrin	U82623
DNA polymerase epsilon	XM_216727
DOC-2; p96 Phosphoprotein	U95177
E1B 19K Bcl-2 binding protein homolog	AF243515
ER81 ETV1	BF524947
ESTAA199109	AA199109/CD372216
ESTAI013714	gb|AI013714
ETF	BF555149
Fibronectin	X15906
FISP-12	NM_022266
GADD153	gb|U36994
GAPDH	NM_017008
Gas-6	D42148
Granulin (GRN)	M97750
guanine nucleotide binding protein G-s	M12673
Gu binding inhibitor of activated STAT1	XM_217188
HB-EGF	L05489
helicase p68 (HUMP68)	AJ010934

Gene name	Accession number
Hic-5	AF314960
Histone H3.3	XM_227461
hNop56	CB567043
ID1	D10862
interferon induced gene	X61381
JAK1 protein tyrosine kinase 1	AJ000556
Jun	X17163
JunB	X54686
JunD	D26307
KS1	U56732
Lot1	U72620
Lysyl oxidase	S77494
Lysyl oxidase-related protein (WS9-14)	AW916312
MAP-kinase phosphatase cpg21	AF013144
megakaryocyte potentiating factor	NM_031658
Mkp1	X84004
Mkp3	X94185
MKP-4	XM_219711
MMP-1 (Collagenase)	M60616
MMP-3 (Stromelysin-1)	X02601
MMP10 (Transin-2)	X05083
Mob-1	U17035
MST2	AJ001529
MUK2	NM_017198
MyoD	M84176
myo-inositol monophosphatase (IMP)	U84038
Nras rat	NM_080766
nucleoside diphosphate kinase puf	gb|M55331
ODC1	NM_012615
p15 (ink4b)	BE126804
P5 protein	X79328
p53	L07909
P-cadherin	AW144786
PDGF	Z14120
PEBP2a1	AB025797
Phosducin-like protein (PhLP)	L15354
poly ADP-ribose glycohydrolase	AB019366
polyhomeotic mRNA	BF555212
PP-1	D00859

Gene name	Accession number
Rap1B GTP-binding protein	U07795
R-esp2	L14463
RhoA	XM_228860
Sex comb on midleg homolog	AW140736
single strand DNA-binding protein	AF121893
Syndecan-1	NM_013026
TDAG51	NM_017180
thrombospondin 1	BE127004
TIMP-2	S72594
Tsc36	NM_024369
USF-2	NM_031139
VD3R	NM_017058
WT1	NM_031534

B Cluster Members

Tables B.1-B.4 list the cluster members of clusters A-D together with their GO terms. GO terms were obtained with the help of the freely available software EASE (http://david.niaid.nih.gov/david/ease.htm), which allows local annotation of gene lists [Hosack et al. 03].
Some genes, whose GO terms could not be retrieved with EASE, were annotated using the Bioinformatic Harvester provided by EMBL (http://harvester.embl.de). If no information could be obtained for a certain gene, the corresponding field is left empty.
It is of note that of the 25 genes in cluster B only 24 could be screened for SRF targets, since the Ensembl-ID for polyhomeotic mRNA was unavailable.

Table B.1: Members of cluster A (37 genes)

Genes in cluster A (abbreviation)	GO Biological Process	GO Cellular Component	GO Molecular Function
aryl hydrocarbon receptor (AHR)	-apoptosis -cell cycle -regulation of transcription -response to stress -signal transduction -xenobiotic metabolism	-cytoplasm -transcription factor complex	ligand-dependent nuclear receptor activity -protein binding transcription factor activity
BCSC-1	negative regulation of cell cycle		tumor suppressor
cAMP dependent protein kinase (prkar2b)	-protein phosphorylation -signal transduction	-cAMP-dependent protein kinase complex	-cAMP binding -cAMP-dependent protein kinase regulator activity

Genes in cluster A (abbreviation)	GO Biological Process	GO Cellular Component	GO Molecular Function
cca1	-tRNA processing	mitochondrion	-ATP binding -tRNA adenylyl-transferase activity magnesium ion binding
c-myc	-cell cycle arrest -iron ion homeostasis -regulation of transcription	nucleus	-protein binding -transcription factor activity
CSF-1	-macrophage differentiation -positive regulation of cell proliferation -hemopoiesis	integral to membrane	macrophage colony stimulating factor receptor binding
cytocentrin	-mitosis -small GTPase mediated signal transduction -transport	membrane	-GTPase activator activity
DNA polymerase epsilon	DNA replication	nucleus	-DNA binding -epsilon DNA polymerase activity -transferase activity
ER81/ETV1			
ESTAI013714			
ETF			
Gu binding inhibitor of activated STAT1	-JAK-STAT cascade -regulation of transcription -signal transduction -ubiquitin cycle	nucleus	-zinc ion binding -DNA binding -transcription corepressor activity -ATP-dependent RNA helicase activity
DOC-2	cell proliferation		
growth arrest specific 6 (Gas6)		extracellular space	apoptosis inhibitor activity

Genes in cluster A (abbreviation)	GO Biological Process	GO Cellular Component	GO Molecular Function
Hic-5	-positive regulation of transcription	intracellular	-androgen receptor binding -transcription co-activator activity -zinc ion binding
Inhibitor of DNA binding 1 (ID1)	-negative regulation of transcription -regulation of angiogenesis	transcription factor complex	-DNA binding -protein binding
Jun	-regulation of cell cycle -regulation of transcription	nuclear chromosome	-transcription factor activity
Lysyl oxidase related protein (WS9-14)	-protein modification -aging -cell adhesion	-extracellular space -membrane	-protein-lysine 6-oxidase activity -scavenger receptor activity -copper ion binding -electron transporter activity
megakaryocyte potentiating factor	cell adhesion	membrane	oxidoreductase activity
Mkp1	-cell cycle -intracellular signaling cascade -protein dephosphorylation	nucleus	MAP kinase phosphatase activity
MST2	-apoptosis -protein phosphorylation -signal transduction	-cytoplasm -protein kinase CK2 complex	-ATP binding -protein kinase CK2 activity -protein-tyrosine kinase activity
MUK2	-JNK cascade -protein phosphorylation	-cytosol -focal adhesion	-ATP binding -MAP kinase kinase kinase activity -protein-tyrosine kinase activity

Genes in cluster A (abbreviation)	GO Biological Process	GO Cellular Component	GO Molecular Function
myogenic differentiation 1 (MyoD)	-cell differentiation -myogenesis -regulation of transcription	-transcription factor complex	-transcription factor activity -enhancer binding -protein binding
myo-inositol monophosphatase (IMP)	-carbohydrate metabolism -phosphate metabolism -phosphatidylinositol biosynthesis -signal transduction	extrachromosomal circular DNA	-inositol-1(or 4)-monophosphatase activity -magnesium ion binding
Nras	-RAS protein signal transduction -regulation of cell cycle	-cytoplasm -membrane fraction -plasma membrane	-ATP binding -GTP binding -RAS small monomeric GTPase activity -protein binding
p15(ink4b)	-regulation of cyclin dependent protein kinase activity -cell cycle arrest -negative regulation of cell proliferation	-nucleus -cytoplasm	-cyclin-dependent protein kinase inhibitor activity
P-cadherin	-protein binding -calcium ion binding	-integral to membrane	-homophilic cell adhesion -sensory perception -visual perception -cell adhesion
PEB2a1			
phosducin-like protein (PhLP)	-electron transport -phototransduction -signal transduction -vision	cytosol	-electron transporter activity -regulator of G-protein signaling activity

Genes in cluster A (abbreviation)	GO Biological Process	GO Cellular Component	GO Molecular Function
poly ADP-ribose glycohydrolase	small GTPase mediated signal transduction	cytoplasm	poly(ADP-ribose) glycohydrolase activity
Sex comb on midleg homolog	-morphogenesis -regulation of transcription	nucleus	-transcription factor activity
sequence-specific single-stranded DNA-binding protein	-regulation of transcription	nucleus	-single-stranded -DNA binding -transcription regulator activity
tissue inhibitor of metallo-proteinase 2 (TIMP-2)		-extracellular matrix -extracellular space	-enzyme activator activity -metalloendo-peptidase inhibitor activity
Tsc36	transport	extracellular space	-calcium ion binding -heparin binding
USF-2	-regulation of transcription	-transcription factor complex	-DNA binding -transcription factor activity -protein binding
vitamin D receptor (VD3R)	-calcium ion homeostasis -induction of apoptosis by hormones -regulation of transcription -skeletal develop ment	-extrachromosomal circular DNA -nuclear chromosome	-steroid hormone receptor activity -transcription factor activity -vitamin D binding -vitamin D3 receptor activity
Wilms tumor 1 (WT1)	-eye morphogenesis -negative regulation of cell cycle -regulation of transcription	transcription factor complex	-protein binding -transcription factor activity

Table B.2: Members of cluster B (25 genes)

Genes in cluster B (abbreviation)	GO Biological Process	GO Cellular Component	GO Molecular Function
alpha actinin	-cell motility -invasive growth	actin cytoskeleton	-GTP binding -actin bundling activity -calcium ion binding -structural constituent of cytoskeleton
antioxydant enzyme AOE372	-anti-apoptosis peroxidase reaction -response to oxidative stress	cytoplasm	-apoptosis inhibitor activity -electron transporter activity -peroxidase activity -selenium binding
CAMK-related peptide	neurogenesis	microtubule associated complex	kinase activity
guanine nucleotide binding protein G-s (Gnas)	-G-protein signaling, adenylate cyclase activating pathway -energy reserve metabolism -response to drug	-Golgi trans cisterna -membrane fraction	-GTP binding -heterotrimeric G-protein GTPase activity -signal transducer activity
helicase p68	Cell growth	Nucleus	Hydrolase activity ATP-dependent helicase activity RNA binding
Histone H3.3	chromosome organization and biogenesis	-nucleus -chromosome	DNA binding
hNop56			
Janus kinase 1 (JAK1)	-intracellular signaling cascade -protein phosphorylation	cytoskeleton	-ATP binding -protein-tyrosine kinase activity

Genes in cluster B (abbreviation)	GO Biological Process	GO Cellular Component	GO Molecular Function
JunB	-regulation of cell cycle -regulation of transcription	-chromatin -nucleus	-transcription co-activator activity -transcription co-repressor activity
JunD	-regulation of cell cycle -regulation of transcription	-chromatin -nucleus	transcription factor activity
KRAB/zinc finger suppressor protein 1 (KS-1)	regulation of transcription	transcription factor complex	nucleic acid binding
MAP-kinase phosphatase (cpg21)	protein dephosphorylation	-cytoplasm -nucleus	-MAP kinase phosphatase activity
matrix metalloproteinase 1 (MMP1)	collagen catabolism	-extracellular space -extracellular matrix	-interstitial collagenase activity -zinc ion binding -calcium ion binding
matrix metalloproteinase 3 (MMP3)	collagen catabolism	extracellular matrix	-calcium ion binding -stromelysin 1 activity zinc ion binding
MKP-3	-apoptosis -inactivation of MAPK -protein dephosphorylation	cytoplasm	-MAP kinase phosphatase activity
MKP-4	-inactivation of MAPK -protein dephosphorylation	-nucleus -cytoplasm	MAP kinase phosphatase activity
nucleoside diphosphate kinase puf	-CTP, GTP, UTP biosynthesis -negative regulation of cell proliferation -regulation of transcription	microtubule	-ATP binding -nucleoside diphosphate kinase activity -transcription factor activity

Genes in cluster B (abbreviation)	GO Biological Process	GO Cellular Component	GO Molecular Function
ornithine decarboxylase 1 (ODC1)	polyamine biosynthesis	mitochondrial inner membrane	ornithine decarboxylase activity
p53	-damage response, signal transduction resulting in induction of apoptosis -negative regulation of cell cycle -cell differentiation -cell aging -nucleotide-excision repair -regulation of mitochondrial membrane permeability -regulation of transcription	-nucleolus -mitochondrion	-transcription factor activity -zinc ion binding -ATP binding -DNA strand annealing activity -copper ion binding -protein binding -nuclease activity
platelet derived growth factor alpha (Pdgfa)	regulation of cell cycle	-extracellular space -membrane	-growth factor activity
polyhomeotic mRNA			
protein phosphatase 1 (PP-1)	-glycogen metabolism -protein dephosphorylation -regulation of protein biosynthesis	-cytoplasm -nucleoplasm	-calcium-dependent protein serine/threonine phosphatase activity
Rap1b	-regulation of cell cycle -small GTPase mediated signal transduction	membrane	-GTP binding -RAS small monomeric GTPase activity

Genes in cluster B (abbreviation)	GO Biological Process	GO Cellular Component	GO Molecular Function
RhoA	-protein transport -Rho protein signal transduction -actin cytoskeleton organization and biogenesis -positive regulation of I-kappaB kinase/ NF-kappaB cascade -positive regulation of NF-kappaB-nucleus import	-cytoskeleton -membrane	-GTP binding -GTPase activity -magnesium ion binding
T-cell death associated gene (TDAG51)	-embryogenesis -morphogenesis	integral to membrane	exo-alpha-sialidase activity

Table B.3: Members of cluster C (12 genes)

Genes in cluster C (abbreviation)	GO Biological Process	GO Cellular Component	GO Molecular Function
E1B 19K Bcl-2 binding protein homolog	anti-apoptosis	-integral to membrane -mitochondrion	-apoptosis activator activity -apoptosis inhibitor activity
ESTAA199109			
GADD153	-cell cycle arrest -regulation of transcription	transcription factor TFIID complex	transcription factor activity
glyceraldehyde-3-phosphate dehydrogenase (GAPDH)	glycolysis	cytoplasm	phosphorylating activity
granulin	lipid catabolism	extracellular space	-calcium ion binding -cytokine activity -phospholipase A2 activity

Genes in cluster C (abbreviation)	GO Biological Process	GO Cellular Component	GO Molecular Function
interferon induced gene			
matrix metalloproteinase 10 (MMP10)	collagen catabolism	extracellular matrix	-stromelysin 2 activity -zinc ion binding
Mob-1	-cell motility -chemotaxis -inflammatory response -muscle development -positive regulation of cell proliferation -protein secretion -signal transduction	extracellular	chemokine activity
Lot1	regulation of cell cycle	transcription factor complex	
protein disulfide isomerase-related protein (P5)	-electron transport -protein folding	-endoplasmic reticulum	-electron transporter activity -protein disulfide isomerase activity
R-esp2	-frizzled signaling pathway -negative regulation of transcription	-heterotrimeric G-protein complex -nucleus	-protein kinase activity -transcription corepressor activity
syndecan 1	histogenesis and organogenesis	integral to membrane	cytoskeletal protein binding

Table B.4: Members of cluster D (8 genes)

Genes in cluster D (abbreviation)	GO Biological Process	GO Cellular Component	GO Molecular Function
actin-related protein 3 homolog (Arp3)	cell motility	Arp2/3 protein complex	-protein binding -structural molecule activity

Genes in cluster D (abbreviation)	GO Biological Process	GO Cellular Component	GO Molecular Function
balbc aldose reductase-related protein	aldehyde metabolism	extracellular space	aldehyde reductase activity
Cox2	-negative regulation of cell proliferation -positive regulation of cell proliferation -prostaglandin biosynthesis -response to oxidative stress	-endoplasmic reticulum -nuclear membrane	-oxidoreductase activity, acting on single donors with incorporation of molecular oxygen -prostaglandin-endoperoxide synthase activity
Fibronectin	-acute-phase response -cell adhesion -wound healing	-extracellular matrix -extracellular space	-cell adhesion molecule activity -heparin binding -oxidoreductase activity
FISP-12	-DNA metabolism -angiogenesis -cell adhesion -intracellular signaling cascade regulation of cell growth	extracellular matrix	-cell adhesion molecule activity -heparin binding -insulin-like growth factor binding
HB-EGF	-EGF receptor signaling pathway -regulation of heart rate	-extracellular space -integral to membrane	-growth factor activity -heparin binding
Lysyl oxidase (LOX)	protein modification	extracellular matrix	-copper ion binding -protein-lysine 6-oxidase activity

Genes in cluster D (abbreviation)	GO Biological Process	GO Cellular Component	GO Molecular Function
thrombospondin-1	-neurogenesis -blood coagulation -cell adhesion -cell motility -development	extracellular region	-protein binding -calcium ion binding -signal transducer activity -structural molecule activity -endopeptidase inhibitor activity -heparin binding

C Ets Matrices

Matrix-ID	Consensus sequence	Information content in nats	Source	Database	Remarks
M00340	KRCAGGAAR TRNKT	9.54	9 compiled binding sequences	TRANS FAC_ PUBLIC	c-Ets-2
M00339	RCAGGAAGTG NNTNS	8.81	21 compiled binding sequences	TRANS FAC_ PUBLIC	c-Ets-1
M00678	YTACTTCCTG	10.2	5 compiled binding sequences	TRANS FAC	Tel-2
M00746	RNWMBAGGA ART	8.93	8 compiled sequences	TRANS FAC	ELF-1
M00658	WGAGGAAG	7.22	14 compiled binding sequences	TRANS FAC	PU.1
M00341	VCCGGAAGN GCR	9.75	12 compiled binding sequences	TRANS FAC_ PUBLIC	GA binding protein
M00108	ACCGGAAGNG	8.63	7 compiled binding sequences from 3 genes	TRANS FAC_ PUBLIC	nuclear respiratory factor 2
M00531	YRNCAGGAAG YRNSTBDS	11.48	6 genomic binding sites	TRANS FAC	new ets-related factor 1a

Matrix-ID	Consensus sequence	Information content in nats	Source	Database	Remarks
M00771	ANNCACTTC CTG	7.24	48 compiled sequences	TRANS FAC	Ets, Q4
M00971	ACTTCCTS	6.04	81 compiled sequences	TRANS FAC	Ets, Q6
M00743	CMGGAAGY	7.76	8 compiled sequences	TRANS FAC	c-Ets-1
M00655	ACWTCCK	6.79	10 compiled binding sequences	TRANS FAC	PEA3
MA0098	NWTCCD	4.82	SELEX	JASPAR	c-ETS
M00025	NNNNCCGGAA RTNN	6.77	31 selected sites	TRANS FAC_ PUBLIC	Elk-1, single binding sites in variable distances to SRF-binding sites
M00032	NCMGGAW GYN	7.07	SELEX	TRANS FAC_ PUBLIC	c-Ets-1(p54)
M00074	NNACMGGA WRTNN	5.96	SELEX	TRANS FAC_ PUBLIC	c-Ets-1(p54)

Matrix-ID	Consensus sequence	Information content in nats	Source	Database	Remarks
M00007	NAAACMGGA AGTNCVH	9.06	binding elements from 4 genes	TRANS FAC_ PUBLIC	Elk-1

Table C.1: The first column lists the matrix-IDs as stored in T-reg. The consensus sequence displayed in the second column gives for each position the most frequent nucleotide (B = C/G/T, D = A/G/T, H = A/C/T, K = G/T, M = A/C, N = any nucleotide, R = A/G, S = G/C, V = A/C/G, W = A/T, Y = C/T). In the third column, the information content of the matrices based on the natural logarithm is shown. The fourth column indicates the data on which the matrix is based, SELEX stands for artificial binding site selection. The sixth column contains the Ets protein which binds to the described Ets-site.

Bibliography

[AMARATUNGA & CABRERA 04] D. Amaratunga and J. Cabrera, *EXPLORATION AND ANALYSIS OF DNA MICROARRAY AND PROTEIN ARRAY DATA*, John Wiley & Sons, 2004.

[BELAGULI ET AL.97] S.B. Belaguli, L.A. Schildmeyer, and R.J. Schwartz, *Organization and Myogenic Restricted Expression of the Murine Serum Response Factor Gene*, THE JOURNAL OF BIOLOGICAL CHEMISTRY **272** (1997), 18222–18231.

[BENJAMINI & HOCHBERG 95] Y. Benjamini and Y. Hochberg, *Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing*, Journal of the Royal Statistical Society B **57** (1995), 289–300.

[BOLSHAKOVA & AZUAJE 03] N Bolshakova and F. Azuaje, *Improving expression data mining through cluster validation*, Proc. of the 4th Annual IEEE Conf. on Information Technology Applications in Biomedicine (2003), 19–22.

[BUCHWALTER ET AL.04] G. Buchwalter, C. Gross, and W. Bohdan, *Ets ternary complex transcription factors*, Gene **324** (2004), 1–14.

[CAMPBELL ET AL.98] S. Campbell, R. Khosravi-Far, K.L. Rossman, G.J. Clark, and C.J. Der, *Increasing complexity of Ras signaling*, Oncogene **15** (1998), 1395–1413.

[CAMPS ET AL.98] M. Camps, A. Nichols, C. Gillieron, B. Antonsson, M. Muda, C. Chabert, U. Boschert, and S. Arkinstall, *Catalytic Activation of the Phosphatase MKP-3 by ERK2 Mitogen-Activated Protein Kinase*, SCIENCE **280** (1998), 1262–1264.

[CAMPS ET AL.00] M. Camps, A. Nichols, and S. Arkinstall, *Dual specificity phosphatases: a gene family for control of MAP kinase function*, The FASEB Journal **14** (2000), 6–16.

[CHO ET AL.98] R.J. Cho, M.J. Campbell, E.A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T.G. Wolfsberg, A.E. Gabrielian, D. Landsman, D.J. Lockhart, and R.W. Davis, *A Genome-Wide Transcriptional Analysis of the Mitotic Cell Cycle*, Molecular Cell **2** (1998), 65–73.

[CURWEN ET AL.04] V. Curwen, E. Eyras, D. Andrews, L. Clarke, E. Mongin, S.M.J. Searle, and M. Clamp, *The Ensembl Automatic Gene Annotation System*, Genome Research **14** (2004), 942–950.

[DIETERICH ET AL.02] C. Dieterich, B. Cusack, H. Wang, K. Rateitschak, A. Krause, and M. Vingron, *Annotating regulatory DNA based on man-mouse genomic comparison*, BIOINFORMATICS **18** (2002), S84–S90.

[DIETERICH ET AL.03] C. Dieterich, R. Herwig, and M. Vingron, *Exploring potential target genes of signaling pathways by predicting conserved transcription factor binding sites*, BIOINFORMATICS **19** (2003), ii50–ii56.

[DIETERICH ET AL.05] C. Dieterich, S. Grossmann, A. Tanzer, S. Roepcke, P.F. Arndt, P.F. Stadler, and M. Vingron, *Comparative promoter region analysis powered by CORG.*, BMC Genomics **6** (2005), 24–34.

[EFRON 92] B. Efron, *SIX QUESTIONS RAISED BY THE BOOTSTRAP*, EXPLORING THE LIMITS OF BOOTSRAP (1992), 99–126.

[EISEN ET AL.98] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein, *Cluster analysis and display of genome-wide expression patterns*, Proc. Natl. Acad. Sci. USA **95** (1998), 14863–14868.

[FURUKAWA ET AL.03] T. Furukawa, M. Sunamura, F. Motoi, S. Matsuno, and A. Horii, *Potential Tumor Suppressive Pathway Involving DUSP6/MKP-3 in Pancreatic Cancer*, American Journal of Pathology **162** (2003), 1807–1815.

[FUTSCHIK & KASABOV 02] M.E. Futschik and N.K. Kasabov, *Fuzzy Clustering of Gene Expression Data*, IEEE Press (2002).

[GINEITIS & TREISMAN 01] D. Gineitis and R. Treisman, *Differential Usage of Signal Transduction Pathways Defines Two Types of Serum Response Factor Target Gene*, THE JOURNAL OF BIOLOGICAL CHEMISTRY **276** (2001), 24531–24539.

[GRAVES & PETERSEN 98] B.J. Graves and J.M. Petersen, *Specificity within the ets Family of Transcription Factors*, Adv Cancer Res. **75** (1998), 1–55.

[GÜNTER & BUNKE 02] S. Günter and H. Bunke, *Validation Indices for Graph Clustering*, Pattern Recognition Letters **24** (2002), 1107–1113.

[HERZEL ET AL.02] H. Herzel, D. Beule, S. Kielbasa, J. Korbel, C. Sers, A. Malik, H. Eickhoff, H. Lehrach, and J. Schuchhardt, *Extracting information from cDNA arrays*, CHAOS **11** (2002), 98–107.

[HOSACK ET AL.03] D.A. Hosack, G.Jr. Dennis, B.T. Sherman, H.C. Lane, and R.A. Lempicki, *Identifying biological themes within lists of genes with EASE*, Genome Biology **4** (2003), R70.1–R.70–8.

[IYER ET AL.99] V.R. Iyer, M.B. Eisen, D.T. Ross, G. Schuler, T. Moore, J.C.F. Lee, J.M. Trent, L.M. Staudt, J.Jr. Hudson, M.S. Boguski, D. Lashkari, D. Shalon, D. Botstein, and P.O. Brown, *The Transcriptional Program in the Response of Human Fibroblasts to Serum*, SCIENCE **283** (1999), 83–87.

[KAROLCHIK ET AL.03] D. Karolchik, R. Baertsch, M. Diekhans, T.S. Furey, A. Hinrichs, Y.T. Lu, K.M. Roskin, M. Schwartz, C.W. Sugnet, D.J. Thomas, R.J. Weber, D. Haussler, and W.J. Kent, *The UCSC Genome Browser Database*, Nucleic Acids Research **31** (2003), 51–54.

[KASZA ET AL.05] A. Kasza, A. O'Donnell, K. Gascoigne, L.A.H. Zeef, A. Hayes, and A.D. Sharrocks, *The ETS Domain Transcription Factor Elk-1 Regulates the Expression of Its Partner Protein, SRF*, THE JOURNAL OF BIOLOGICAL CHEMISTRY **280** (2005), 1149–1155.

[KERR ET AL.00] K. Kerr, M. Martin, and G.A. Churchill, *Analysis of Variance for Gene Expression Microarray Data*, J. Comp. Biol. **7** (2000), 819–837.

[KIM ET AL.05] T.H. Kim, L.O. Barrera, M. Zheng, C. Qu, M.A. Singer, T.A. Richmond, Y. Wu, R.D. Green, and B. Ren, *A high-resolution map of active promoters in the human genome*, Nature (2005).

[LIN ET AL.92] X. Lin, Z. Wang, L. Gu, and T.F. Deuel, *Functional Analysis of the Human Platelet-derived Growth Factor A-chain Promoter Region*, THE JOURNAL OF BIOLOGICAL CHEMISTRY **267** (1992), 25614–25619.

[MIANO 03] J.M. Miano, *Serum response factor: toggling between disparate programs of gene expression*, Journal of Molecular and Cellular Cardiology **35** (2003), 577–593.

[MO ET AL.01] Y. Mo, W. Ho, K. Johnston, and R. Marmorstein, *Crystal Structure of a Ternary SAP-1/SRF/c-fos SRE DNA Complex*, J. Mol. Biol. **314** (2001), 495–506.

[MÜLLER 04] B. Müller, *Expressionsanalyse von Zielgenen der RAS-Onkogen abhängigen Signalübertragung in einem definierten Zeitfenster nach Aktivierung des RAS-Onkogens*, 2004.

[PÉRIER ET AL.00] R.C. Périer, V. Paz, T. Junier, C. Bonnard, and P. Bucher, *The Eukaryotic Promoter Database (EPD)*, Nucleic Acids Research **28** (2000), 302–303.

[PHILIPPAR ET AL.04] U. Philippar, G. Schratt, C. Dieterich, J.M. Müller, P. Galóczy, F.B. Engel, M.T. Keating, F. Gertler, R. Schüle, M. Vingron, and A. Nordheim, *The SRF Target Gene Fhl2 Antagonizes RhoA/MAL-Dependent Activation of SRF*, Molecular Cell **16** (2004), 876–880.

[POLLARD & VAN DER LAAN 05] K.S. Pollard and M.J. van der Laan, *Cluster Analysis of Genomic Data with Applications in R*, U.C. Berkeley Division of Biostatistics Working Paper Series (2005).

[RAHMANN ET AL.03] S. Rahmann, T. Müller, and M. Vingron, *On the Power of Profiles for Transcription Factor Binding Site Detection*, Statistical Applications in Genetics and Molecular Biology **2** (2003).

[ROEPCKE ET AL.05] S. Roepcke, S. Grossmann, S. Rahmann, and M. Vingron, *T-Reg Comparator: an analysis tool for the comparison of position weight matrices*, Nucleic Acids Research **33** (2005), W438–W441.

[SCHULZE ET AL.01] A. Schulze, K. Lehmann, HB. Jefferies, M. Mcmahon, and J. Downward, *Analysis of the transcriptional program induced by Raf in epithelial cells.*, Genes & Development **15** (2001), 981–994.

[SERS ET AL.02] C. Sers, O.I. Tchernitsa, J. Zuber, L. Diatchenko, B. Zhumabayeva, S. Desai, S. Htun, K. Hyder, K. Wiechen, A. Agoulnik, K.M. Scharff, P.D. Siebert, and R. Schäfer, *Gene expression profiling in RAS oncogene-transformed cell lines and in solid tumors using subtractive suppression hybridization and cDNA arrays*, Advan. Enzyme Regul. **42** (2002), 63–82.

[SPENCER ET AL.99] J.A. Spencer, M.J. Major, and R.P. Misra, *Basic Fibroblast Growth Factor Activates Serum Response Factor Gene Expression by Multiple Distinct Signaling Mechanisms*, MOLECULAR AND CELLULAR BIOLOGY **19** (1999), 3977–3988.

[SUZUKI ET AL.01] Y. Suzuki, H. Taira, T. Tsunoda, J. Mizushima-Sugano, J. Sese, H. Hata, T. Ota, T. Isogai, T. Tanaka, S. Morishita, K. Okubo, Y. Sakaki, Y. Nakamura, A. Suyama, and S. Sugano, *Diverse transcriptional initiation revealed by fine, large-scale mapping of mRNA start sites*, EMBO Rep **2** (2001), 388–393.

[SUZUKI ET AL.04] Y. Suzuki, R. Yamashita, S. Sugano, and K. Nakai, *DBTSS, DataBase of Transcriptional Start Sites: progress report 2004*, Nucleic Acids Research **32** (2004), D78–D81.

[TAMAYO ET AL.99] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E.S. Lander, and T.R. Golub, *Interpreting patterns of gene expression* *with self-organizing maps: Methods and application to hematopoietic differentiation*, Proc. Natl. Acad. Sci. USA **96** (1999), 2907–2912.

[TULLAI ET AL.04] J.W. Tullai, M.E. Schaffer, S. Mullenbrock, S. Kasif, and G.M. Cooper, *Identification of Transcription Factor Binding Sites Upstream of Human Genes Regulated by the Phosphatidylinositol 3-Kinase and MEK/ERK Signaling Pathways*, THE JOURNAL OF BIOLOGICAL CHEMISTRY **279** (2004), 20167–20177.

[VASSEUR ET AL.03] S. Vasseur, C. Malicet, E.L. Calvo, C. Labrie, P. Berthezene, C.J. Dagorn, and J.L. Iovanna, *Gene expression profiling by DNA microarray analysis in mouse embryonic fibroblasts transformed by rasV12 mutated protein and the E1A oncogene*, Molecular Cancer (2003), 2–19.

[VICKERS ET AL.04] E.R. Vickers, A. Kasza, I.A. Kurnaz, A. Seifert, L.A.H. Zeef, A. O'Donnell, A. Hayes, and A.D. Sharrocks, *Ternary Complex Factor-Serum Response Factor Complex-Regulated Gene Activity Is Required for Cellular Proliferation and Inhibition of Apoptotic Cell Death*, MOLECULAR AND CELLULAR BIOLOGY **24** (2004), 10340–10351.

[VOLMAT ET AL.01] V. Volmat, C. Montserrat, S. Arkinstall, J. Pouysségur, and P. Lenormand, *The nucleus, a site for signal termination by sequestration and inactivation of p42/p44 MAP kinases*, Journal of Cell Science **114** (2001), 3433–3443.

[WASSERMAN & SANDELIN 04] W.W. Wasserman and A. Sandelin, *APPLIED BIOINFORMATICS FOR THE IDENTIFICATION OF REGULATORY ELEMENTS*, NATURE REVIEWS GENETICS **5** (2004), 276–287.

[ZUBER ET AL.00] J. Zuber, O.I. Tchernitsa, B. Hinzmann, A. Schmitz, M. Grips, M. Hellriegel, C. Sers, A. Rosenthal, and R. Schäfer, *A genome-wide survey of RAS transformation targets*, nature genetics **24** (2000), 144–152.