5 Applications for CORG

This chapter focuses on the analysis and interpretation of microarray data in conjunction with CORG. Microarray data constitute a read out of the transcriptional program. A major challenge in the area of functional genomics is to correlate gene expression with promoter level events in higher organisms. Ideally, one would like to link expression changes to underlying signaling pathways that activate specific transcription factors. Binding sites of the corresponding transcription factors should be present in co-expressed genes. Binding sites can be predicted by searching for recurring motifs in upstream regions of co-expressed genes. This approach has been successfully applied in yeast where the intergenic regions are small enough to inspect them for binding sites (e.g. Tavazoie et al. (1999); Bussemaker et al. (2001)).

However, this approach cannot be directly translated to mammalian genomes where the position of the promoter region is often unknown and intergenic regions tend to be large. This results in a prohibitively high number of false positives in binding site prediction. We address this problem by focusing binding site prediction on genomic segments that are conserved between man and mouse as in the CORG database.

Two applications will be presented in this chapter:

1. **Detection of putative cell cycle regulators.** Genes were grouped into clusters of co-expressed genes. These clusters correspond to cell cycle phases. The resulting distributions of conserved predicted binding sites were statistically assessed.

2. **Investigation of SRF-mediated gene expression.** We studied SRF-induced genes from different experimental contexts. Detailed promoter and comparative analyses of the induced gene sets were carried out.

5.1 Binding site distributions across cell cycle phases

We exploit the distribution of evolutionarily conserved, predicted binding sites over different groups of co-expressed genes as an indicator for functionality of the predicted binding sites. The rationale is that when a factor plays a role in the co-expression of a group of genes, we ought to observe these functional binding sites sticking out from the random occurrences of predicted binding sites. A deviation from the random

distribution for a particular factor should thus indicate a functional role for these binding sites.

We will exemplify this for human cell cycle data, with the genes that peak in a particular phase of the cell cycle taking the role of the co-expressed group. Gene expression data are taken from Whitfield et al. (2002) who studied expression levels of genes in cycling HeLa cells. Based on the expression levels, they identified genes that are periodically upregulated and assigned each of them to one out of five expression clusters corresponding to the cell cycle phases G1/S, S, G2, G2/M and M/G1.

Figure 5.1 condenses our knowledge on the cell cycle into a single graphic (see legend text). The duration of each cell cycle phase is depicted as its perimeter proportion of the whole cycle. The duration of the whole cell cycle varies greatly from one cell type to another. Fly embryos have the shortest known cell cycles, each lasting as little as 8 minutes, while the cell cycle of a mammalian liver cell can last longer than a year. Whitfield et al. (2002) monitored the cell cycle of HeLa S3 cells that divide on average every 12 hours.

Evolutionarily conserved, predicted TFBSs and the cell cycle phase assignment of genes according to Whitfield et al. (2002) can be combined to search transcription factor candidates that may play a role in cell cycle regulation. Both kinds of data can be represented in matrix form. Thus, we denote the matrix representing predicted TFBSs for m motifs in upstream regions of n genes as B with dimensions m, n and the matrix assigning k periodically expressed genes to l cell cycle phases as C with dimensions k, l. Since B holds all genes in CORG ($\approx$ 12,000) and C contains only a partially overlapping subset (i.e. $n \neq k$), both matrices need to be adjusted. All gene entries $G \in B \cap C$ are pulled out of both matrices to yield B' and C'. The product A of B' and C' is defined by $a_{ik} = \sum_j b'_{ij} c'_{jk}$ with i as TFBS index, j as gene index and k as phase index. Thus, rows of matrix A contain count distributions of PWM hits over all cell cycle phase. Given this approach, we now discuss all settings to obtain B and C.

5.1.1 Binding site prediction

The 120 employed PWMs were retrieved from the TRANSFAC database release 7.1 (Matys et al., 2003). Only PWMs with both sensitivity and specificity quality value above 0.9 according to an ROC curve as define by Rahmann et al. (2003) were selected. The corresponding PWM identifiers are listed at `http://corg.molgen.mpg.de/cellcycle/matrix_ids.txt`. The proportion of false negative and false positive observations was computed with respect to a signal and a background model. We set the false negative level to 10% in our initial searches. The number of false positive hits was cut by an upper bound on the accepted p-value ($p < 0.0005$). In case of alternative translational start sites, all unique binding sites for a given gene were pooled.

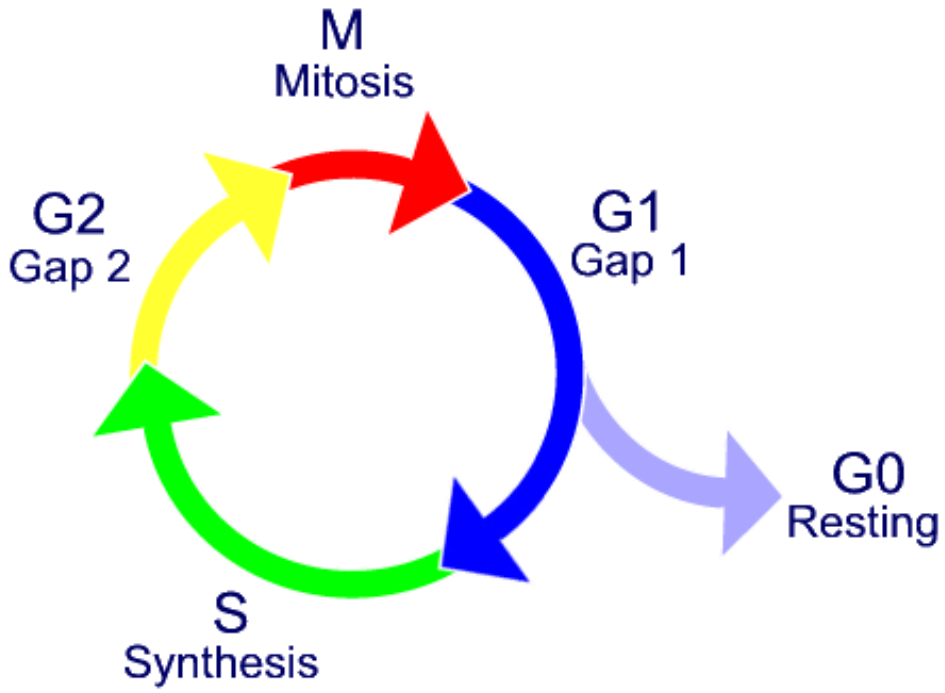


Figure 5.1: Mitotic cell cycle. The cell cycle is a sequence of recurring events from one cell division to the next. The following cell cycle phases can be distinguished: The *G1 phase* is the first growth phase. Cells might exit the cell cycle at this stage and remain in *G0 phase*. During G0, cells maintain a quiescent state. G1 is followed by the *S phase*, during which the DNA is replicated. *G2 phase* is the second growth phase. Cells prepare now for the *M phase* or mitosis and cytokinesis, the actual division of the cell into two daughter cells. The cell cycle stops at several checkpoints and can only proceed if certain conditions are met, for example, if the cell has reached a certain diameter. Some cells, such as neurons, never divide again once they become locked in G0 phase. Image source: http://www.med.unibs.it/~marchesi/dna.html

For this data set, we could unambiguously identify an overlap of 592 genes with the CORG database.

5.1.2 Association mining

Bias in PWM counts across cell cycle phases PWM count distributions that deviate significantly from the null distribution might indicate a preferential association of particular binding sites in certain cell cycle phases. A reasonable choice for the null distribution $Q = (q_i)$ is given by dividing the summed length of the conserved sequence in upstream regions of genes associated to phase i by the total length of con-

	Cell cycle phases				
Sequence length in bp	G1/S	S	G2	G2.M	M.G1
5' UTR	25132	17875	27160	34982	18671
Non-exonic sequence	104614	67871	129803	162927	75845
Total sequence	129746	85746	156963	197909	94516

Table 5.1: Length of conserved sequence per expression cluster. The length of conserved sequence (in bp) across the three upstream sequence categories and cell cycle phases. The null probability distributions are based on these data.

served sequence. The respective sequence lengths are shown in Table 5.1 for exonic and non-exonic segments in the corresponding upstream regions. The overall length of conserved sequence per expression cluster constitutes the search space of binding site predictions. Random occurrences of binding sites would be uniformly distributed across the search space, whereas functional sites should show some cluster preference in their distribution pattern.

Next, an exact test is applied to determine which of the observed PWM-phase distributions obey the assumed background distribution Q. Additionally, a correction for multiple testing is crucial since one p-value is computed for each PWM. By employing the concept of False Discovery Rates (Storey and Tibshirani, 2003), we adjust the size of our result set such that we expect at most one false discovery. This analysis is carried out for the non-exonic upstream regions and for 5'-UTRs separately. As a final result we obtain a number of transcription factors that we predict to play a functional role in the progression of the cell cycle and compare these results in the light of the experimental literature.

Test for deviation from background distribution An exact p-value for observed counts (n_i) is calculated with an exact likelihood ratio test as follows. We use the generalized likelihood ratio statistic

$$\mathcal{L} := 2 \cdot \sum_{i=1}^{k} n_i \cdot \log(n_i/(nq_i)), \tag{5.1}$$

where n_i is the observed number of samples in category i, $n = \sum n_i$ is the sample size and q_i is the null probability of category i. The null hypothesis H_0 is that the counts (n_i) arise from sampling n times a random category (cell cycle phase) from the null distribution, i.e., that (n_i) has a multinomial distribution with sample size n and category distribution Q. The p-value of an observed value ℓ of $\mathcal{L}$ is defined as the probability $p(\ell) = \mathbb{P}_{H_0}(\mathcal{L} \geq \ell)$ that the value of the test statistic $\mathcal{L}$ exceeds the observed value ℓ when the counts are in fact generated by sampling from the null distribution. The test statistic $\mathcal{L}$ is a relative entropy-like measure.

Recently Bejerano (2003) and Rahmann (2003) have shown that it is possible to efficiently compute exact p-values in this situation; for this study, we used the MATLAB implementation provided by Rahmann.

Alternatively, we can obtain approximate p-values using an asymptotic χ^2 approximation. Here, the test statistic

$$\chi^2 := \sum_{i=1}^{k} (n_i - nq_i)^2/(nq_i) \tag{5.2}$$

has an approximate χ^2 distribution with $k - 1 = 4$ degrees of freedom given $k = 5$ cell cycle phases. Again, the hypothesis of whether the sample was drawn from Q is tested. The asymptotic approximation alternative should be preferred in case of large sample sizes. The computation time for approximate p-values is only a tiny fraction of the running time of the exact method. Additionally, accuracy of the approximated p-values increases with sample size.

Correcting for multiple testing Storey and Tibshirani (2003) proposed a more liberal criterion to correct for multiple testing than traditional methods like the Bonferroni correction. The principal idea is to use the concept of the False Discovery Rate (FDR, Benjamini and Hochberg (1995)).

Definition 5.1. The False discovery rate is the expected proportion of false rejections given a rejection rule $\mathcal{R}$, e.g. "Reject null hypothesis H_i if the corresponding p-value P_i meets some condition c." More formally,
$\mathrm{FDR}(\mathcal{R}) = E\{\text{proportion of rejected } H_i \text{ that are actually true}\}$.

Benjamini and Hochberg (1995) also provide a useful algorithm to keep the FDR under a preset level α. Let

$$i_\epsilon = \texttt{argmax}_i \left\{ P_i \leq \frac{i\alpha}{np_0} \right\} \tag{5.3}$$

with p_0 being the proportion of true H_i. Then the FDR is always lower than α, if all $P_i \leq P_{i_\alpha}$ are rejected.

However, p_0 is usually not known. The most conservative choice would be to set p_0 to 1, which is the well-known Bonferroni correction. Storey and Tibshirani (2003) offer a way to estimate p_0 from the observed P_i. Exploiting the fact that null P_i values are uniformly distributed, a reasonable estimate can be formed.

$$\hat{p_0}(\lambda) = \frac{\#[P_i > \lambda]}{n(1-\lambda)} \tag{5.4}$$

A tuning parameter λ was introduced to yield estimates of p_0 for different p-value thresholds. Storey and Tibshirani (2003) try to approximate $\hat{p_0}(\lambda = 1)$ by fitting a natural cubic spline to data points for a range of $\lambda \in [0, 0.95]$. Alternatively, a bootstrapping approach (sampling with replacement) is employed to find the optimal λ (see Storey (2002) for details).

Once an estimate ($\hat{p_0}$) is found, the FDR for some significance threshold t is estimated by

$$\widehat{\mathrm{FDR}}(t) = \frac{\hat{p_0} nt}{\#\left[P_i \leq t\right]} \tag{5.5}$$

Our observed uncorrected p-values from both tests were submitted to the R routines available at `http://genomine.org/qvalue`. Since we have a small number of p-values, we employed the recommended "bootstrap" method to estimate π_0. The FDR level was tuned such that we would expect one false prediction on average.

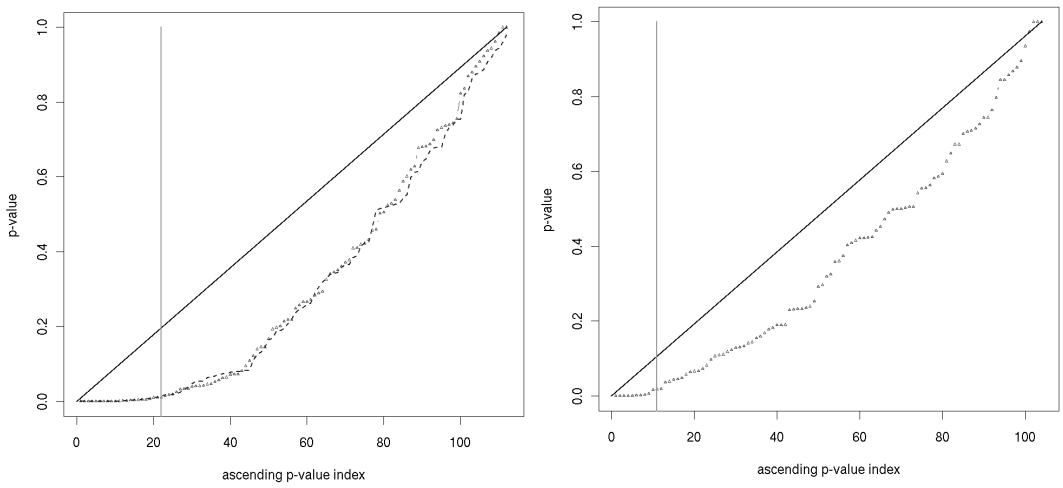


Figure 5.2: ***p*-values for conserved predicted binding site distributions.**
Left (Non-exonic binding site distributions): All 112 *p*-values are plotted in ascending order. Exact *p*-values are plotted in dark grey and the χ^2 approximated *p*-values in light grey dots. The black line depicts the random case where all *p*-values are uniformly distributed. The vertical line marks the highest accepted *p*-value. All in all 22 *p*-values were accepted.
Right (Exonic binding site distributions): All 104 *p*-values are plotted in ascending order. Exact *p*-values are plotted in dark grey. The black line depicts the random case where all *p*-values are uniformly distributed. The vertical line marks the highest accepted *p*-value. All in all 11 *p*-values were accepted.

5.1.3 Significant deviant binding site distributions

A cell-cycle relevant subset of CNBs from CORG (Dieterich et al., 2003b) has been scanned for matches to 120 selected matrix representations (PWMs) of known TFBSs as contained in the TRANSFAC database (Matys et al., 2003). We predicted 24,375 binding sites being part of a man-mouse conserved sequence segment. 3,838 of our predicted binding sites fall into exonic regions. Intriguingly, significantly fewer motif hits were predicted per nucleotide in 5'-UTR sequences (Binomial test, p-value $< 10^{-16}$).

No hits were reported for 8 matrices as none of the few hits were conserved. Figure 5.2a shows the sorted p-values of the exact test and the approximate test of all TFBS phase distributions in non-exonic conserved sequence. Random p-values are uniformly distributed and would ascend along the diagonal (black line).

We were interested whether the ranking between both sets of p-values (exact and approximate) is preserved. Using Kendall's rank correlation test with Kendall's coefficient $\tau = 0.939$, we obtained a p-value of $< 10^{-16}$ for the null hypothesis that both lists are not correlated ($\tau = 0$).

Figure 5.2b depicts the results for conserved predicted TFBSs in 5'UTR regions. 104 exact p-values are plotted in ascending order. No p-value approximation was done due to the small counts of TFBSs.

Correcting for multiple testing As shown in Section 5.1.2, an assessment of the proportion of true null p-values (p_0) is crucial to the concept of the false discovery rate.

Our estimates for p_0 are 0.584 and 0.5 for non-exonic and 5'UTR TFBSs, respectivly. The lowest attainable FDR is 0.04 for non-exonic and 0.07 for 5'UTR TFBS distributions.

We chose to adjust the number of accepted significant observations to the level of one expected false discovery. This means to set a FDR threshold of 0.045 for the non-exonic PWM phase distributions and 0.096 for the 5'-UTR PWM phase distributions. By applying these thresholds, we accept 22 and 11 PWMs for the two data sets, respectively. The accepted PWMs are presented in Table 5.2 and 5.3.

Figure 5.3a summarizes all PWM phase distributions for non-exonic binding sites that appear in Table 5.2. Similarly, the content of Table 5.3 is visualized in Figure 5.3b. TFBS counts were normalized by subtracting the distribution mean and subsequent division by the standard deviation. The normalized values (Z-scores) were then translated into grey scale colors (see Figure legend).

Transfac motif	Cell cycle phases					p-value
	G1.S	S	G2	G2.M	M.G1	
V.NERF.Q2	168	67	99	199	90	9.210769e-08
V.PAX9.B	152	49	101	148	65	1.071894e-06
V.OLF1.01	120	44	66	134	63	2.0846165e-06
V.COUP.01	107	67	97	228	78	5.1196418e-06
V.STAF.02	148	53	101	148	79	1.5220174e-05
V.TEL2.Q6	245	154	216	370	197	4.9835619e-05
V.ARP1.01	138	57	104	200	91	0.00010639304
V.T3R.01	30	11	15	45	33	0.00010802171
V.PTF1BETA.Q6	165	100	125	240	110	0.00014531276
V.IK3.01	94	40	75	160	64	0.00037316641
V.RREB1.01	26	6	13	13	3	0.00073057134
V.COUP.DR1.Q6	107	57	76	154	74	0.0010877052
V.PPARG.03	84	32	70	138	57	0.0016508134
V.HOX13.01	14	6	3	25	10	0.0027401688
V.MAF.Q6	114	59	85	153	81	0.0030366968
V.AP4.01	197	85	215	247	146	0.0038111711
V.APOLYA.B	4	4	6	25	3	0.0043054911
V.FXR.Q3	88	46	78	159	66	0.0050114715
V.EVI1.06	44	45	67	123	47	0.0063601175
V.FOXO3.01	22	17	14	48	19	0.009670219
V.EVI1.02	49	35	53	115	40	0.010321056
V.MEF3.B	137	55	121	186	83	0.01089912
Null distribution	0.19335	0.12544	0.23991	0.30113	0.14018	

Table 5.2: Top 22 non-random distributions of TRANSFAC motifs for non-exonic sequence. 1 out of the 22 listed motifs is expected to be a false discovery (FDR level of 0.045).

Transfac motif	Cell cycle phases					p-value
	G1.S	S	G2	G2.M	M.G1	
V.NERF.Q2	14	19	77	60	15	2.4700964e-11
V.OLF1.01	5	10	35	44	11	1.4235621e-06
V.HEN1.01	0	1	11	2	0	5.9439633e-05
V.AR.Q2	9	0	15	22	3	0.00017224456
V.TAL1BETAE47.01	0	3	10	10	0	0.00033778459
V.RREB1.01	0	1	0	2	6	0.0018728352
V.IK3.01	8	18	13	18	2	0.0019495133
V.MAF.Q6	10	15	27	47	15	0.0032300874
V.PAX9.B	21	13	38	58	17	0.0056357379
V.FREAC3.01	2	6	2	9	0	0.016426396
V.NFKAPPAB65.01	16	10	30	35	7	0.018244817
Null distribution	0.20297	0.14436	0.21935	0.28252	0.15079	

Table 5.3: Top 11 non-random distributions of TRANSFAC motifs for exonic sequence. 1 false discovery is expected on average in the set of the 11 listed motifs (FDR level of 0.096).

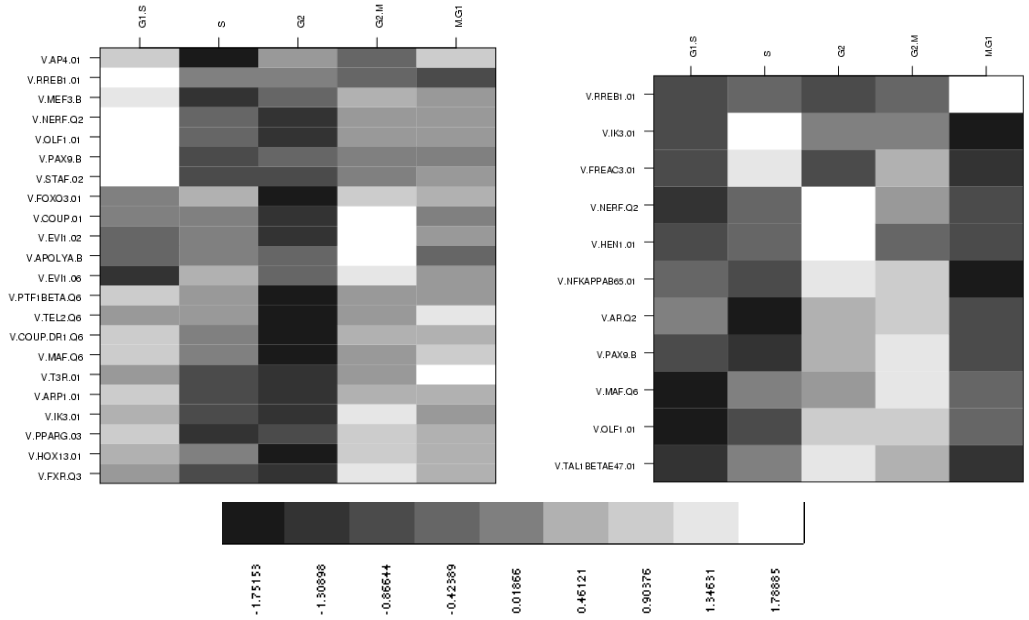


Figure 5.3: Normalized association matrices. *Left*: All TFBS distributions that were selected as truly significant are shown for non-exonic upstream regions. Matrix entries are divided by their column sum and subsequently normalized with the following formula: $(Observed - \mu)/\sigma$ where μ is the row mean value and σ the standard deviation of the row distribution. *Right*: The same for 5' UTR upstream regions.

5.1.4 Biological implications

Many conserved predicted binding sites are not uniformly distributed across cell cycle phases. We will now discuss potential links of our findings to cell cycle regulation for PWMs shown in Figure 5.3.

Starting with TFBS distributions that peak in G1/S, we initially consider a set of 6 PWMs (V$NERF_Q2, V$MEF3_B, V$OLF1_01, V$RREB1_01, V$PAX9_B, V$STAF_02). Not all of the corresponding transcription factors have already been studied in the context of cell proliferation or cell cycle. NERF is an antagonist of ELF-1. Both are Ets transcription factors that play an important role in blood vessel development (Gaspar et al., 2002). Zhang et al. (2003) demonstrated the repressive action of transiently expressed RREB (ras-responsive element binding protein 1) on the promoter of p16 alias *Cdkn2a*, which is a CDK inhibitory protein. Other factors (PAX9, OLF1 and its alternative splice variant EBF) have been linked to developmental processes.

Moving on to the S phase, we recognize a generally lower relative number of predicted TFBS. V$EVI1_Q6 and V$FOXO3_01 have an almost equal relative amount of binding sites in the S and G2/M phase.

Cappellini et al. (2003) established a causal relationship of FOXO1, FOXO3 and cell cycle regulation. They showed that a constitutively activated PI3K/Akt axis triggers phosphorylation of FOXO1 and FOXO3. Thus, FOXO1 and FOXO3 were mainly found in the cytoplasm and did not induce gene expression of *p27-KIP1 (Cdkn1b)* causing cell cycle arrest.

Several examples of enriched G2/M motifs were discovered (e.g. V$PPARG_03 and V_FXR_Q3). V$PPARG_03 denotes elements which are bound by the peroxisome proliferator-activated receptor γ. Wang et al. (2001) could inhibit S-phase entry and expression of the cyclin D1-dependent serine-threonine kinase (Cdk) by natural and synthetic ligands of PPARγ.

Another member of the thyroid hormone/retinoid receptor subfamily of nuclear hormone receptors is the thyroid hormone receptor (T3Rα/β). It has its highest number of relative binding sites in M/G1 phase. T3R also affects cell cycle regulation and alters expression levels of the *Mdm2* oncogene (Qi et al., 1999).

Differences in TFBS distribution for non-exonic and 5'UTR regions. The ranking and the members of the two result sets for non-exonic and 5'UTR TFBS motifs differ. Three motifs (V$FREAC3_01, V$NFKAPPAB65_01, V$AR_Q2) emerge as interesting new candidates in the set of 5'UTR regions. Most notably, the androgen receptor has been reported to regulate the expression of the cyclin-dependent kinase inhibitor 1A (CDKN1A, p21, Cip1) as we would expect from our predictions (Lu et al., 1999). Predicted sites for V$FREAC3_01 peak in S phase. The corresponding transcription factor FOXC1 has been shown to affect a number of developmental processes. Mutations in the *Foxc1* gene have been attributed to ocular, meningeal, cardiac, skeletal and renal anomalies (Lehmann et al., 2003).

All in all, functional TFBS are known to occur in 5'UTR regions of genes and exert their control on transcription there. However, it is difficult to assess their function and contrast it to non-exonic regions.

5.2 Promoter analysis of SRF responsive genes

Serum Response Factor (SRF), a MADS-box transcription factor, regulates the expression of immediate-early genes (IEGs[1]), genes encoding several components of the actin cytoskeleton, and cell-type specific genes, e.g. smooth, cardiac and skeletal muscle or neuronal-specific genes (Miano, 2003; Treisman, 1995). Mouse embryos lacking SRF die before gastrulation and do not form any detectable mesoderm (Arsenian et al.,

[1]The term "immediate-early" denotes rapidly induced genes whose mRNA levels rise sharply within 30 minutes after serum induction.

1998; Weinhold et al., 2000). SRF mediates transcriptional activation by binding to CArG box sequences (Consensus pattern: $\texttt{CC(AT)}_6\texttt{GG}$) in target gene promoters and by recruiting different co-factors. SRF regulates transcription downstream of MAPK signaling in association with ternary complex factors (TCFs) (for a review see Buchwalter et al. (2004)). TCFs bind to ets binding sites present adjacent to CArG boxes in many SRF target gene promoters.

This section summarizes the results of a long-standing collaboration on SRF targets with the group of Alfred Nordheim at the university of Tübingen. We are especially grateful for all the experimental work that was mainly done by Ulrike Phillipar and Gerhart Schratt.

In this study, a whole-genome micro array approach helps to identify new SRF target genes using overexpression of SRF-VP16 in SRF-deficient embryonic stem (ES) cells. We hypothesized that SRF overexpression in ES cells would activate genes that are important for several cellular differentiation pathways, including muscle differentiation. Among others, we identified the gene encoding the LIM-only protein FHL2 as a novel SRF target gene. FHL2 is upregulated in an SRF-dependent manner during differentiation of ES cells and in response to RhoA activation. In addition, our collaborators demonstrated that SRF and FHL2 interact in vitro and in vivo after RhoA activation and form a complex at the promoters of a subset of SRF target genes. We also show that FHL2 acts as a repressor of SRF-induced transcription and specifically represses the MAL-induced activation of smooth muscle gene promoters. We propose that RhoA, SRF, and FHL2 are part of a novel feedback-loop to regulate the expression of SRF target genes during early cardiac muscle differentiation.

5.2.1 Identification of SRF target genes

Expression profiling to identify SRF-regulated genes in ES cells Despite increasing evidence for a role of SRF in various biological processes, further insight into SRF function is hampered by the lack of a comprehensive list of SRF target genes. We decided to use microarray analysis to monitor gene expression as a function of SRF activity at the whole-genome level. We first used transient overexpression of a constitutive active SRF fusion protein, SRF-VP16, in SRF-deficient ES cells. This approach has been shown to lead to a sensitized, robust induction of known SRF target genes, such as *Egr-1* (Schratt et al., 2002). Using SRF-VP16, we expected to identify the majority of SRF target genes independent of activating extracellular cues. SRF-ΔM-VP16, a mutant defective in DNA binding, was used as control. To monitor gene expression profiles of cells transfected with SRF-VP16 or SRF-ΔM-VP16, mRNA from two independent transfections was hybridized to Affymetrix microarrays. Two independent *Srf(-/-)* ES cell lines (81-/-, 100-/-) (Weinhold et al., 2000) were used to control for cell-based variations. We considered those genes as regulated by SRF whose expression was at least threefold higher in each of the samples derived from

SRF-VP16 transfected cells compared to the SRF-ΔM-VP16 control transfected. We obtained a set of 86 genes whose expression was reproducibly induced by SRF-VP16 (Table 5.4). 55 of these genes could be unambiguously mapped to LocusLink identifiers and were further analyzed. A substantial subset of SRF-regulated genes is reported to function in different muscle lineages. Other identified SRF-regulated genes play a role in cell cycle/apoptosis and cytoskeletal organization, wound healing and cellular metabolism. Our analysis recovered 17 previously known SRF target genes, including *Egr-1*, *Sma* (smooth muscle actin) and Vinculin. However, regulation by SRF had not been demonstrated previously for several other genes, e.g. the muscle-specific Lim protein FHL2 or the actin-binding protein Tuftelin-1. Interestingly, our analysis also identified a subset of SRF-regulated genes with a known function in wound healing and angiogenesis, for example *CTGF* (chondrocyte tissue growth factor), *Keratin 17* and *Endothelin-1*. This suggests a role for SRF in these biological processes. Taken together, our microarray approach using the *Srf(-/-)* ES cell system confirmed SRF-regulation for 17 known SRF targets and identified a putative role for SRF in the transcriptional regulation of 69 additional genes.

Table 5.4: SRF induced genes from microarray experiment. Fold activation for two cell lines and two measurements are shown.

Acc.No	Gene name	Fold activation			
Musclespecific (11)		ES81 (1)	ES81(2)	ES100(1)	ES100(2)
Z68618	SM22α	235.1	151.2	251.4	151.4
X13297	αActin (smooth)	98.6	11.8	104.3	137.8
M12347	ACTA1 (skeletal)	80.1	73.6	201.9	88.3
D16497	NPPB	31.8	63.0	71.5	54.0
D88793	CRP1	75.6	32.3	31.2	28.4
M15501	ACTC (cardiac)	9.0	26.3	8.5	14.5
M29793	cTnC	3.4	20.1	14.6	9.8
X14961	HFABP	24.6	4.5	7.6	3.7
AF055889	FHL2	14.7	10.6	3.9	10.6
U28932	CNN1	5.4	4.8	12.8	7.4
AI842649	MLCC	3.6	7.5	20.5	5.4
Cell cycle/apoptosis (10)					
X81584	IGFBP6	144.1	169.9	271.8	185.4
X71922	IGF2	90.8	151.5	112.7	73.1
M28845	EGR1	23.9	56.6	173.4	56.6
AF058798	1433β	17.0	36.4	7.6	12.5
U09268	PAC1	4.0	26.6	11.5	28.9
M35523	CRABP2	15.6	20.4	14.5	11.7
M24377	EGR2	6.8	16.6	10.5	24.2
U20735	JunB	16.9	22.1	1.9	2.3
X67644	IER3	3.2	18.8	4.4	5.2
Z38110	PMP22	15.7	4.1	6.1	3.0
Wound healing/angiogenesis (7)					
M26071	Tissue factor	28.5	104.4	26.5	33.9
M70642	CTGF	38.7	13.5	58.6	5.2
U35233	Endothelin1	15.8	22.6	36.9	18.6
X69619	InhibinbetaA	20.5	15.8	11.3	19.8
M32490	Cyr61	17.0	16.6	19.6	10.3
AF100777	WISP1	7.3	6.2	11.0	3.9
X62700	MuPAR1	7.5	5.4	3.3	3.1
Cytoskeleton (25)					
M69260	Lipocortin 1	101.0	66.8	55.8	81.3
M13805	Keratin 17	17.8	77	104.2	68.4
M14044	Annexin A2	179.0	34.8	19.9	15.6
M36120	Keratin 19	13.5	25.4	121.0	20.3
M22832	Keratin 18	22.5	36.0	54.2	18.6
AI604345	NICE1	11.3	46.1	13.0	58.8
AJ001633	Annexin III	26.0	69.3	13.2	17.0
AI462105	Vinculin	70.4	8.5	10.9	7.3
M77174	Perlecan	3.6	47.3	7.8	4.6

J04953	Gelsolin	15.4	12.9	24.0	7.5
U82624	APP	25.4	10.9	14.3	8.1
AA755126	Keratin 7	4.9	9.7	25.8	14.1
AA600542	Desmoplakin	23.9	7.6	15.2	4.4
X15662	Keratin 8	14.1	16.2	12.0	8.1
AF087824	Claudin6	12.0	20.0	10.6	7.0
AI835858	Tropomyosin 4	23.6	5.3	11.1	5.7
AW060401	TC10	5.8	25.6	10.9	3.1
M15832	Procollagen, type IV	16.3	7.9	3.5	3.7
AB000713	CPE receptor	7.7	6.5	6.1	3.8
AF047704	Tuftelin	4.4	5.6	6.4	4.4
AW046449	Fyn	4.8	4.7	4.6	4.5
M21495	γActin (smooth)	4.8	3.1	3.1	4.4
X15986	Betagalactoside-specific lectin	7.2	7.2	5.8	4.6
Metabolism (9)					
M31775	Cytochrome b558	5.6	11.2	7.3	4.8
J02652	Malic enzyme	6.8	6.4	7.0	4.5
U25739	YSPL1	3.5	7.0	9.4	10.4
X7380	Secretin	4.0	7.4	8.0	5.8
AB025408	Sid478p	7.3	4.3	5.2	5.0
AA726364	Lipoprotein lipase	6.2	16.4	7.6	4.0
X66449	Calcyclin	8.4	14.0	3.8	6.7
M65270	CathepsinB	5.8	4.2	3.7	3.3
Others (15)					
X56602	Interferon induced 15 kDa protein	25.3	11.1	93.6	21.2
M18070	PRNP (prion)	27.6	23.3	10.2	9.6
Y07836	Stra13/Clast5	27.7	27.9	6.0	4.3
X01838	Beta2 microglobulin	32.5	7.5	15.0	3.8
AW125478	IGFBP5 protease	23.4	17.9	4.0	4.7
M61007	C/EBPβ	12.5	8.1	16.9	2.4
AW061260	Nestin	11.1	7.2	10.0	4.3
AI840413	PSD95	5.6	9.4	7.9	6.9
AI849587	Calcium channel 8	6.9	7.5	6.3	3.1
AI842665	TIP1	3.1	5.7	7.9	5.6
AI845915	PolI transcription related factor	7.3	6.0	3.6	3.8
U59807	Cystatin	6.4	5.1	3.9	3.3
AB026569	MSSP	5.0	5.0	3.8	4.9
AI844520	IFI30	3.4	6.5	3.5	3.7
J05261	Protective protein for galactoside	4.2	5.3	3.0	3.1

Identification of conserved SRF binding sites using a comparative genomic approach Our results so far do not address whether SRF regulates genes within the set of 86 genes identified by microarray by directly binding to the respective promoter regions. We employed a mouse/human comparative genomics approach to screen the regulatory regions of the identified 55 genes for putative SRF consensus binding sites. 10 Kb of DNA sequence upstream of the first partially translated exon in the genomic sequences of mouse and human was analyzed. To narrow down the search space, we considered only binding sites that localized within conserved sequence elements of mouse and human genomic sequences (CNBs). Cross-species sequence conservation often correlates with functional importance and helps to reduce the false positive rate. Within the CNBs, we screened for putative SRF binding sites, allowing one base pair exchange from the canonical SRF consensus sequence ($CC(A/T)_6GG$). Such sites often constitute functional SRF binding sites. We identified a total of 21 conserved putative SRF binding sites within the upstream region of the reduced set of 55 genes induced by SRF-VP16 (Table 5.5). We found SRF binding sites in CNBs of 10 (out of 17) previously known SRF targets, including *SM22α, Egr-1* and muscle actins. This illustrates that functional SRF binding sites can be identified with the comparative genomics approach chosen here. The failure to identify the SRF sites within the remaining seven known SRF targets may be explained by the fact that these binding sites are present in enhancer elements further upstream or downstream of the search space (as is for example true for the *CRP-1* gene). In addition, functional SRF binding sites are not always conserved in terms of sequence and genomic location. Eleven previously unrecognized, new SRF target genes were identified with our approach. These can be categorized based on the degree of conservation of the upstream CArG sequence(s) (Table 5.5). Upstream regions of the Tuftelin, FHL2 and Keratin 17 genes contain a canonical CArG sequence that is entirely conserved between the mouse and human genome. CArG sequences in the eight other genes are either conserved but have one basepair exchange (e.g. *Endothelin-1*), or are only partially conserved (e.g. *CTGF*). Taken together, using comparative sequence analysis, we identified SRF binding sites in 10 known and 11 previously uncharacterized SRF target genes. Together with our results from expression profiling, we propose that a significant fraction of genes that contain evolutionary conserved SRF binding sites within their upstream regulatory regions may be directly regulated by SRF.

Rating of in silico binding site predictions To get an idea on the performance of our *in silico* predictions, we evaluated the overlap of our predictions with an *in vivo* large-scale experiment. Ren et al. (2002) monitored the promoter occupancy of 1,200 genes with E2F family members 1 and 4. These 1,200 genes are expressed during cell cycle entry ($G_0 \rightarrow G_1$ transition) in primary fibroblasts.

Table 5.6 provides an overview on the intersection of this data set with our "standard" binding site predictions in the CORG database. Further details are given in the

Table 5.5: Conserved putative SRF binding sites. The last column denotes whether the corresponding binding site has been previously reported as functional in literature.

Acc.No	Gene	CArG sequence	Pos.Human	Pos.Mouse	functional?
M12347	ACTA1 (skeletal)	CCAAATATGG	1066	1141	x
M15501	ACTC (cardiac)	CCAAATAAGG	840	881	x
		CCTTTTAAGG	5796	6325	x
U20735	JunB	CCTAATATGG	1814	1729	x
M24377	Egr2	CCTTTTTTGG	697	790	x
		CCATATATGG	398	414	x
M28845	Egr1	CCATATAAGG	691	679	x
		CCTTATTTGG	653	641	x
		CCATATTAGG	378	377	x
		CCATATATGG	366	364	x
M32490	Cyr61	CCAAATATGG	2283	2099	x
Z68618	SM22α	CCAAATATGG	3793	4276	x
		CCATAAAAGG	3920	4399	x
AF047704	Tuftelin1	CCTTTTAAGG	664	642	
AF055889	FHL2	CCTTATATGG	12584	10751	
M13805	KRT17	CCTATAAAGG	11926	12171	
Conserved/CArGlike					
U28932	CNN1	CTATAAATGG	2794	3214	x
X14961	HFABP	CCTATTTCGG	145	121	x
M29793	cTnC	CCATACAAGG	538	511	x
		CTAATTTTGG	524	497	x
U35233	Endothelin1	CTATATTTGG	9007	10062	
		ACATAAAAGG	700	715	
		CCTTAAGTGG	3241	2948	
X67644	IER3	CCTAACTTGG	8415	8055	
M77174	Perlecan	CCCTATATGG	13742	12704	
Z38110	PMP22	CCGTTAAAGG	1847	1664	
U09268	PAC1	CCTTGTATGG	140	134	

table's annotation. We cannot distinguish E2F-1 from E2F-4 binding as both proteins have the same DNA binding domain and thus binding site pattern. Consequently, the number of predicted E2F sites remains constant for both factors. Despite this shortcoming, both data sets show an impressive intersection. This is far from being random as expressed in the tiny p-values for both overlaps. From this result, we conclude that our binding site predictions are meaningful since they are supported by biological evidence.

TF	experiment	this work	Overlap	P
E2F-4	79	240	43	6.1×10^{-8}
E2F-1,4	38	240	26	6.3×10^{-8}

Table 5.6: In this table we compare the number of E2F target genes with the biological binding data of Ren et al. (2002). The first row corresponds to targets of E2F-4 and the second row are targets of both E2F-1 and E2F-4. The second column denotes the experimentally observed number of bound promoter regions followed by the number of conserved promoters which contain a known E2F binding motif. The last two columns give the overlap with experiment and the corresponding p-value calculated from the hypergeometric distribution. The total overlap of our conserved promoter regions with the experimental data set is 886.

Enrichment of predicted SRF sites in induced gene set Further support for the biological significance of our predictions comes from a statistical argument: Do we see an enrichment of *in silico* predicted binding sites in our set of SRF induced genes ?

Relaxed CArG box Our default definition of a conserved CArG box was a pair of aligned sequences that stem from the consensus CC[AT](6)GG allowing one mismatch in each sequence. For example, the two aligned strings CCATGAATGG (man) and CCATCAATGG (mouse) comply with this definition. According to this definition, 21 out of 86 mouse genes have at least one CArG box, which is a proportion of 24.4%. On a genome-wide level, 1864 out of 13540 mouse genes that show conservation to man do also have one or more CArG boxes. This is equivalent to a proportion of 13.8%. The two proportions differ significantly with a *p*-value of 0.006 (Exact binomial test, 21 successes and 86 trials, p=0.138)

Stringent CArG box A more stringent definition of a conserved CArG box is a pair of aligned sequences that stem from the consensus CC[AT](6)GG and are identical. Here, the two aligned strings CCATTAATGG (man) and CCATTAATGG (mouse) are an example of this stringent definition. With this stringent definition, 9 out of 86 mouse genes have at least one CArG box, which is a proportion of 10.5%. On a genome-wide level, 151 out of 13540 mouse genes that show conservation to man do also have one or more CArG boxes. This is equivalent to a

proportion of 1.1%. Here, the difference of proportion is even more pronounced. An exact binomial test with 9 successes in 86 trials with an expected frequency of p=0.011 yields a p-value of 5×10^{-7}.

5.2.2 Experimental Validation of SRF-regulated genes by RT-PCR and ChIP

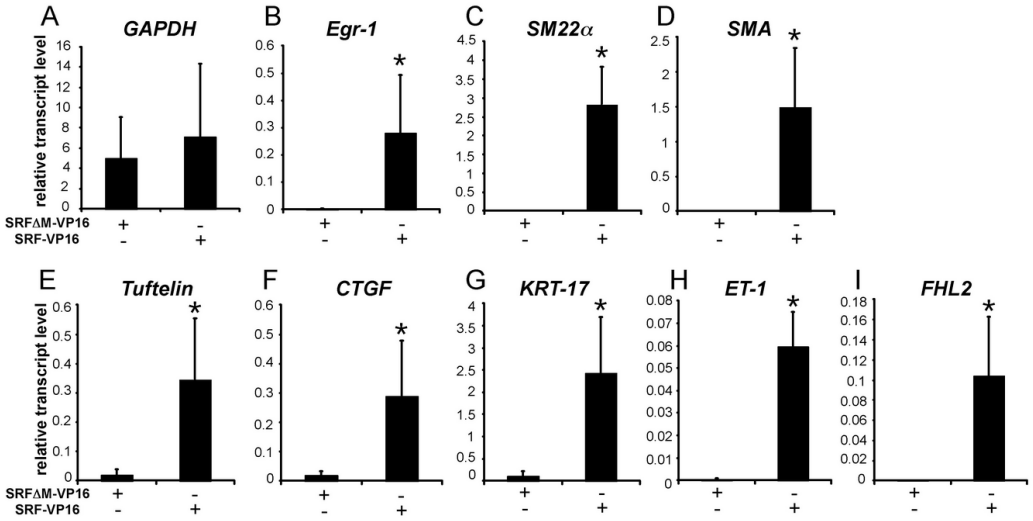


Figure 5.4: Validation of selected SRF targets by RT-PCR. Validation of selected microarray targets, Quantitative RT-PCR analysis was performed with mRNA from Srf(-/-) ES cells (cell line 100) transiently transfected with SRF-VP16 or SRF(M-VP16, using specific primers for *GAPDH* (A), *Egr-1* (B). *SM22α* (C), *SMA* (D), *Tuftelin* (E), *CTGF* (F), *KRT-17* (G), *ET-1* (H) and *FHL2* (I). Values represent transcript levels relative to the endogenous house-keeping gene *HPRT* and are the mean of three independent experiments.

Having identified potential SRF-regulated genes by combined microarray expression analysis and genomic bioinformatics, we verified both SRF-dependent expression and in vivo SRF promoter binding for several gene candidate genes, using quantitative RT-PCR and chromatin immunoprecipitation assays (ChIP), respectively. For this analysis, we focused on candidate genes which represented different functional classes of putative SRF targets: *Tuftelin, CTGF, Keratin-17, Endothelin-1 (ET-1)* and *Fhl2*. For quantitative RT-PCR analysis, undifferentiated 100 *Srf(-/-)* ES cells were transiently transfected with SRF-VP16 or SRFΔM-VP16 expression vectors. In an *Srf(-/-)* background, the mRNA levels of the known SRF target genes *Egr-1, SM22α* and smooth muscle actin (*SMA*) were increased about 160-, 5000- and 15 000-fold, respectively, by SRF-VP16, as compared with control transfected cells. This confirms the capacity of SRF-VP16 to induce the expression of SRF target genes independently of

the activity of upstream signaling cascades (Figure 5.4 B-D). In contrast, expression of the house-keeping gene *GAPDH* did not change upon expression of SRF-VP16, confirming the gene-specificity of the SRF-VP16 protein (Figure 5.4 A). Tuftelin mRNA levels were induced 38-fold, *CTGF* mRNA levels 19-fold, $Keratin-17$ mRNA levels 30-fold, $ET-1$ mRNA levels 100-fold and *Fhl2* mRNA levels about 380-fold by SRF-VP16 (Figure 5.4 E-I). Similar results were obtained using an independent *Srf(-/-)* ES cell line (data not shown). Thus, we were able to validate by quantitative RT-PCR that SRF-VP16 induced the expression of five of the candidate genes originally identified in the above microarray study.

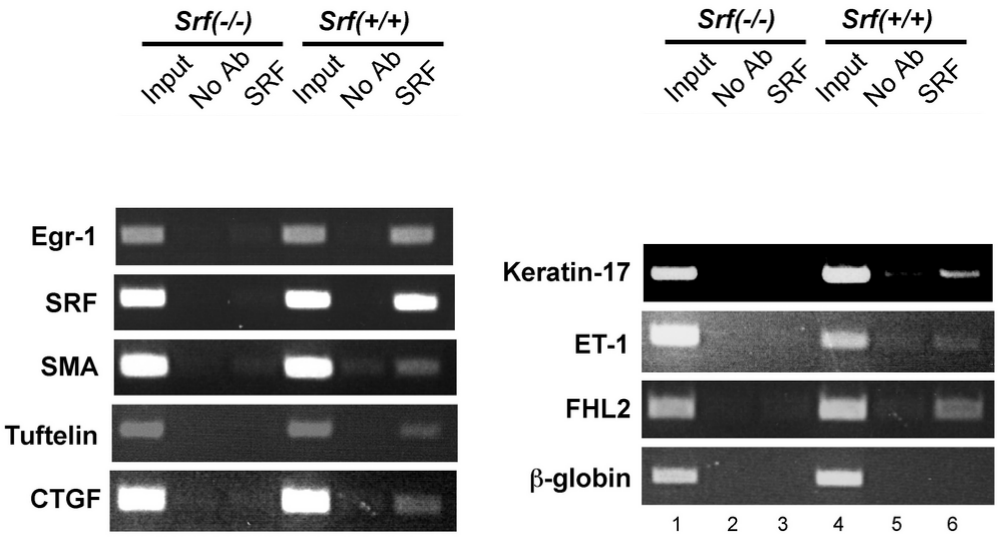


Figure 5.5: In-vivo SRF target validation by ChIP. SRF binding to the murine promoters of selected microarray targets in vivo. Chromatin of day 8 differentiated 100Srf(-/-) and WT ES cells was immunoprecipitated using an SRF antiserum (lanes 3 and 6) and bound DNA fragments were amplified by PCR using primers specific for the CArG containing promoter regions of the indicated genes. Incubation without antibody was used as control (lanes 2 and 5). 1% of the isolated genomic DNA served as input (lanes 1 and 4).

Endogenous SRF binds the murine promoters of Tuftelin, CTGF, Keratin-17, Endothelin-1 and *Fhl2* in vivo To determine whether SRF is directly bound to the identified CArG Box sequences in the promoter regions of *Tuftelin, CTGF, Keratin-17, Endothelin-1* and *Fhl2* in vivo, we performed ChIP assays. Differentiated ES cells were used for the ChIP assays expecting that some SRF target promoters, e.g. muscle-specific promoters, might not be occupied by SRF in undifferentiated stem cells (Manabe and Owens, 2001). 100 Srf(-/-) and WT14.1 ES cells were differentiated for eight days under monolayer conditions in the absence of LIF and the chromatin was immunoprecipitated using an anti-SRF antiserum (Figure 5.5). SRF bound specif-

ically to the CArG box sequences of the known SRF target genes Egr-1, Srf and SMA in differentiated WT ES cells (lane 6). No signal was observed in differentiated 100 Srf(-/-) ES cells (lane 3) and where antibody was absent (Figure 5.5, lanes 2 and 5). SRF also bound specifically to the identified CArG box sequences in the $Tuftelin, CTGF, Keratin-17, ET-1$ and $FHL2$ promoters in differentiated WT ES cells (Figure 5.5, lane 6). The promoter region of the β-globin gene, which contains no CArG Box sequence, did not show any signal after immunoprecipitation with the SRF antibody, further demonstrating the specificity of our results. Taken together, these results demonstrate that the promoter regions of Tuftelin, CTGF, Keratin-17, Endothelin-1, and Fhl2 are bound by SRF in native chromatin of differentiating ES cells, thereby identifying these genes as novel direct SRF target genes.

5.3 Comparison to the LPS response of dendritic cells.

Another induction process where SRF is known to participate is the response of human dendritic cells to LPS treatment. LPS (lipopolysaccharide) is a component of the cell wall of gram-negative bacteria like *E. coli* and is detected by the cell via CD14 and the TLR4 receptor (Guha and Mackman, 2001a). Firstly, I studied the gene set and promoter constitution of the induced genes. Then, I contrasted our previously defined set of SRF-driven genes with the immediate-early genes of the LPS response. Here, SRF is thought to function in conjunction with Ets proteins (Buchwalter et al., 2004) as part of MAPK signalling. However, one has to be aware of the fact that the underlying data were recorded in two completely different systems: differentiated murine ES cells and human dendritic cells.

5.3.1 Studying the LPS response of dendritic cells

The temporal classification of expression profiles was taken from the original research paper (Huang et al., 2001). The immediate-early genes are of particular interest as they are most likely affected by the initial signal. A subset of these genes is targeted by SRF (see Figure 5.6, left part).

109 genes of the immediate-early class could be assigned to LocusLink identifiers. We found conserved exact matches to known binding site patterns taken from TRANSFAC in the upstream regions of 54 of these genes. The resulting binding site matrix (BS matrix) was compiled from the predicted binding sites of 42 distinct TFs in 54 genes.

Of course, we were curious whether we could rediscover the principal signaling pathways outlined in Guha and Mackman (2001a). They summarize the available experimental knowledge as follows (Figure 5.6): The LPS response is mainly triggered by the TLR4 receptor. As a downstream event, various pathways target 5 transcription

factor complexes (Elk1/SRF, c-Jun/c-Fos, c-Jun/ATF-2(CRE-BP1), CREB/ATF-1 and p50/p65), which bind to 4 different types of binding sites (SRE, AP-1, CRE and NF-κB). In our analysis we attempt to show which groups of genes are targets of these transcription factors.

We rearranged the 42 $\times$ 54 BS matrix to identify dense subgraphs of regulators and genes. Figure 5.7 summarizes our results. Genes with upstream AP-1 and CRE elements cluster together, SRE and NF-κB elements are found in two distinct gene groups. Furthermore, binding sites for SRF and NF-κB do not occur together whereas single genes belonging to these groups may additionally contain AP-1 elements.

Other co-occuring transcription factors are ETF/Sp1 in 10 genes and HNF-3A, C/EBP, GATA-1 in 5 genes. Binding of C/EBPβ alias NF-IL6 to promoters of cytokines is a common phenomenon (Holloway et al., 2001). C/EBPβ has been shown to synergize with NF-κB on the promoters of *IL-8* and *ICAM1*. The conjunction of the latter three binding sites has so far only been reported for the *IL12* promoter in the context of the immune response (Becker et al., 2001).

In Dieterich et al. (2003a), we give an account of these findings and stress the distinct role of SRF in the context of the LPS response, which is supported by two alternative methods for binding site prediction.

5.3.2 Comparison of target gene sets

For a direct comparison of the human LPS vs. the murine ES data set, we have to reiterate our pre-conditions. We could unambigously map 55 genes to LocusLink identifiers for the ES cell set and 109 for the LPS induction. Our in-silico predictions classify 24 genes (out of 55) as direct SRF targets for the ES cell set and 27 genes (out of 109) for the LPS induction experiment. Note that these SRF binding site predictions are based on the conserved CArG box consensus sequence allowing one mismatch as in Section 5.2.1. That is why, we detect herein 24 target genes as opposed to 8 with exact matches in Dieterich et al. (2003a).

The two target gene sets (murine ES cells and human dendritic cells) show a marginal overlap of two orthologs: *JUNB* and *IER3*. IER3 functions in the protection of cells from Fas- or tumor necrosis factor type alpha-induced apoptosis (Wu et al., 1998) and is thought to be induced by NF-κB. JUNB is a potent transcriptional activator and a known target of SRF . Taking a closer look at the binding site positions in both promoters, we observe Ets and SRF sites in close proximity in each promoter (Figures 5.8 and 5.9). These sites may well be functional but experimental evidence is required to clarify the situation. Perez-Albuerne et al. (1993) report on the impact of deleting a downstream CArG box of the JUNB gene, but did not pay much attention to the upstream region. No similar reports exist for the IER3 gene.

Table 5.7: Comparison of target gene sets. HUGO symbols for each gene set member are shown. Two genes (IER3 and JUNB) are shared between both sets and are framed.

LPS response in dendritic cells	ATF4, BIRC3, CCL4, CSF1, CXCL2, DTR, DUSP1, DUSP2, EBI2, EMD, H2AFZ, **IER3**, IL6, **JUNB**, LCP2, MAP3K8, MCL1, NDP, NFKBIA, NR4A3, PIM1, PTP4A1, PTX3, TJP2, TNIP1, TRAF1, WTAP
SRF induction in ES cells	ACTA1, ACTA2, ACTC, CNN1, CTGF, EDN1, EGR1, EGR2, F3, FHL2, GADD45B, GM2A, HSPG2, **IER3**, IGFBP6, INHBA, **JUNB**, KRT8, PMP22, RBMS1, SPARC, TAGLN, TES, TUFT1

We can only speculate on the reason of this result. As this is an active area of research, we just started getting insights on modes of SRF-dependent gene regulation. One explanation of different SRF target sets prevails and gains support from experimental data: The existence of SRF co-factors that convey specificity in gene expression. The myocardin family is such a prime case. When the founding member, Myocardin, is expressed ectopically in nonmuscle cells, it can induce smooth muscle differentiation by its association with serum response factor (Du et al., 2003). Wang et al. (2004) showed that Erk-1/2 mediated growth signals repress myogenic gene expression by replacing myocarding with Elk-1, an Ets-Box transcription factor (see also Figure 5.6).

Thus, SRF alone or in conjunction with various co-factors induces different gene sets as seen in this section.

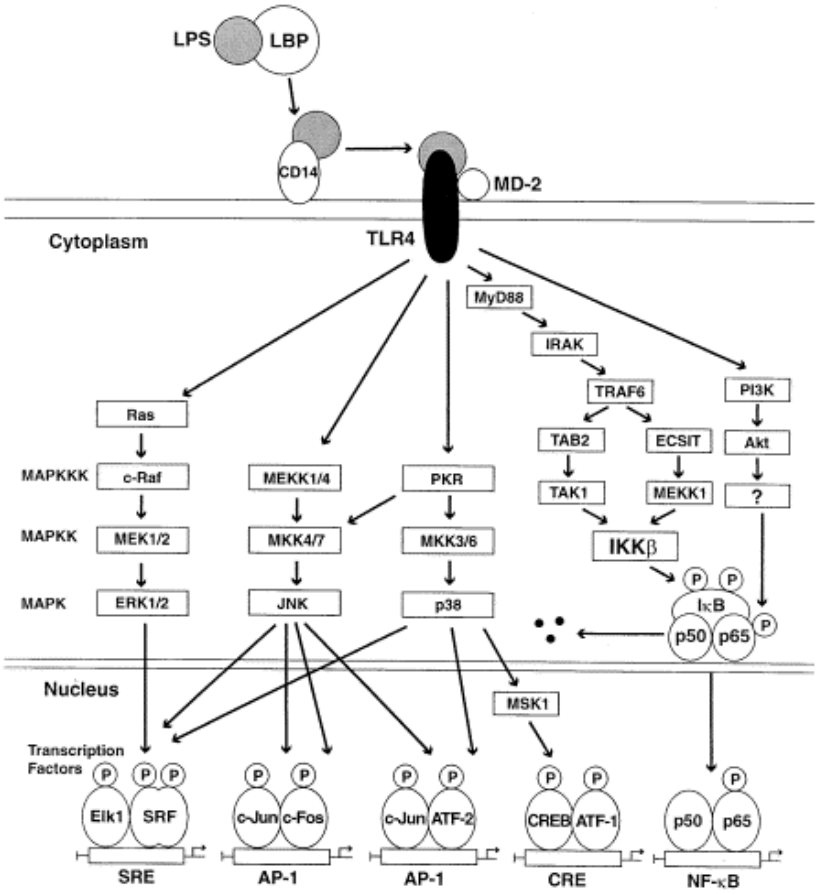
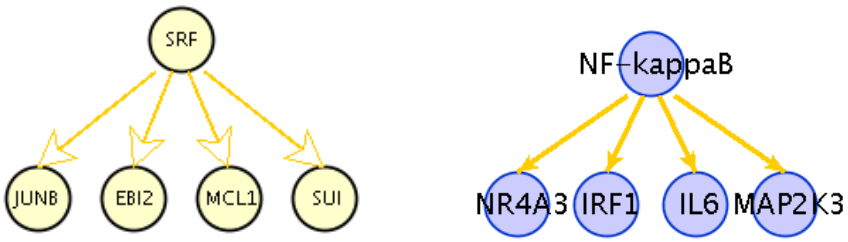


Figure 5.6: Schematic overview of signalling via TLR4 receptor. Bacterial LPS is recognized by the cell via LBP, a soluble protein, which transfers LPS to CD14 and ultimately to the principal cell surface receptor TLR4. Various MAP kinase cascades (Erk 1/2,JNK and p38), the NF-κB pathway and the Akt pathway are triggered by this signal. This signal traverses into the nucleus and activates a set of 5 transcription factor complexes.



FOS, JUN, CREB targets: CD83, DTR, DUSP1, DUSP2, EMD, EXT1, G0S2, GCH1, HSPA1A, IER3, IFIT1, IFIT4, JUNB, MAP2K3, NDP, NR3C1, NR4A3, PTPN1, RELA, SCYA3, SLA

Figure 5.7: Target predictions for the 5 principal transcription factor complexes. This network graph depicts protein-DNA interactions (directed yellow edges) based on conserved exact matches to TRANSFAC site sequences. The following transcription factors are represented: SRF, FOS family (AP-1, c-Fos and Fra-1), JUN family (AP-1, c-Jun, JunB and JunD), CREB and NF-κB

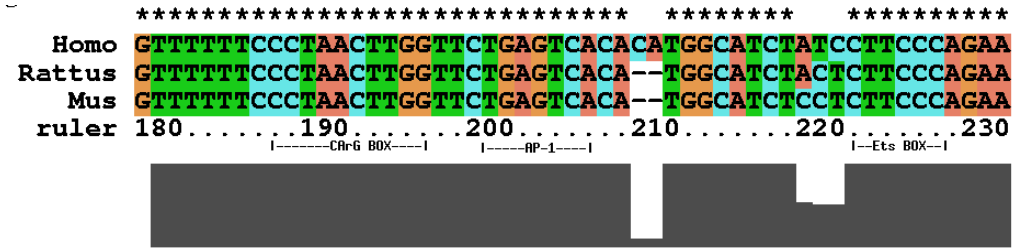


Figure 5.8: Multiple local alignment of promoter region of human IER3 from -8,090 to -7,990 relative to translation start with orthologous mouse and rat sequence. The putative binding sites (SRF, AP-1 and c-Ets) are underlined and are 100% conserved.

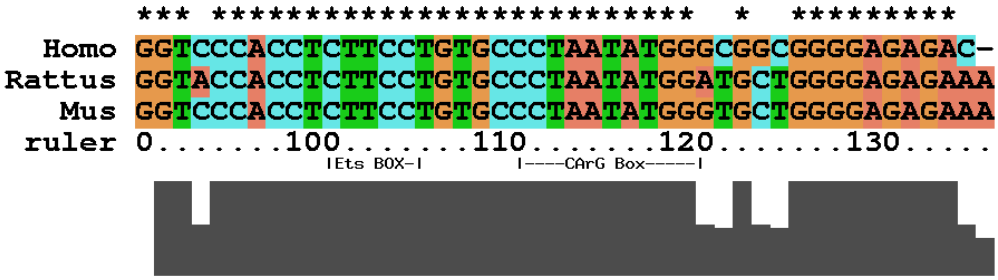


Figure 5.9: Multiple local alignment of promoter region of human JUNB from -6,880 to -6,790 relative to translation start with orthologous mouse and rat sequence. The putative binding sites (SRF and c-Ets) are underlined and are 100% conserved.