CHAPTER 5

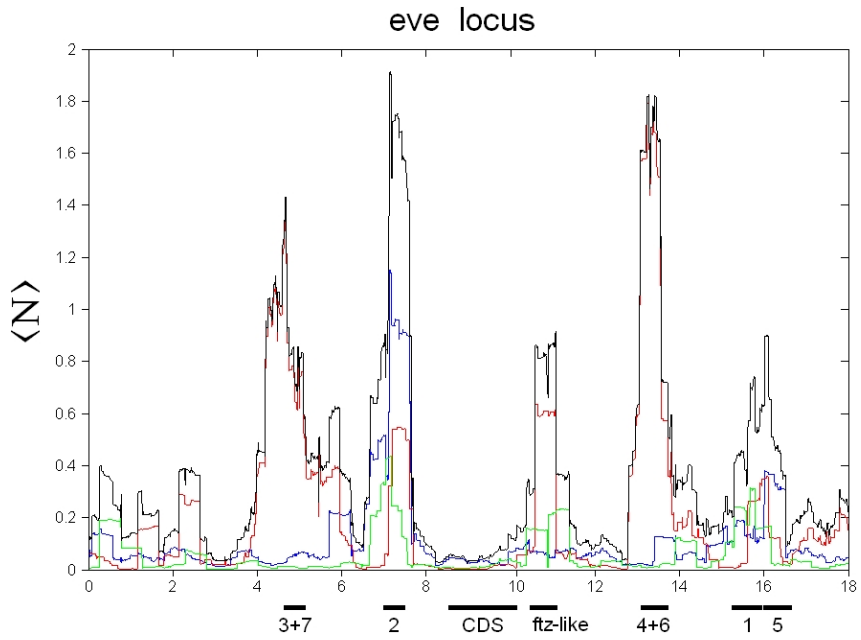
Application of TRAP to higher eukaryotes

The previous Chapter showed the derivation of the TRAP model and its applicability TF binding in yeast. TRAP thereby showed higher predictive power over experimental binding data than state of the art hit based methods as well as alternative biophysical approaches. While these results were encouraging the question arises how well these finding carry over to more complex eukaryotes with more complicated gene structure. The aim of this Chapter is therefore to demonstrate the applicability of TRAP to higher eukaryotic genomes. The focus will thereby lie on three model cases, the *Drosophila eve* promoter, experimental ChIP-PET binding data for the factor P53 and lastly the prediction of target gene for the transcription factor SRF.

5.1 The Drosophila *eve* promoter

To assess whether TRAP, using the generic parameter setting from the yeast analysis, has the potential to be applied to other organisms I first tested how well the method can detect enhancer elements in the *Drosophila melanogaster* gene *eve*. The *eve* gene is expressed in a pattern of seven evenly spaced stripes in the syncytial blastoderms of the developing fly embryo (Small et al., 1993). The expression of each stripe is controlled by a corresponding enhancer element and the presence of several TFs including, Krueppel, Hunchback, Bicoid and Giant (Small et al., 1991). For the purpose of enhancer detection a 500 base pair long window is shifted across the genomic sequence and for each start position *i* the affinity $\langle N \rangle_{window}$ is calculate for the sequence covered by the window. The result of plotting $\langle N \rangle_{window}$ against *i* is shown in Figure 5.1 for the available TRANSFAC matrices BCD_01 (Bicoid), KR_01 (Krueppel) and HB_01 (Hunchback). As illustrated, all experimentally identified enhancer regions can be well detected by this approach indicating that TRAP can be used also in a genetic background other than yeast.

Figure 5.1 – Identification of enhancer stripes in the eve gene

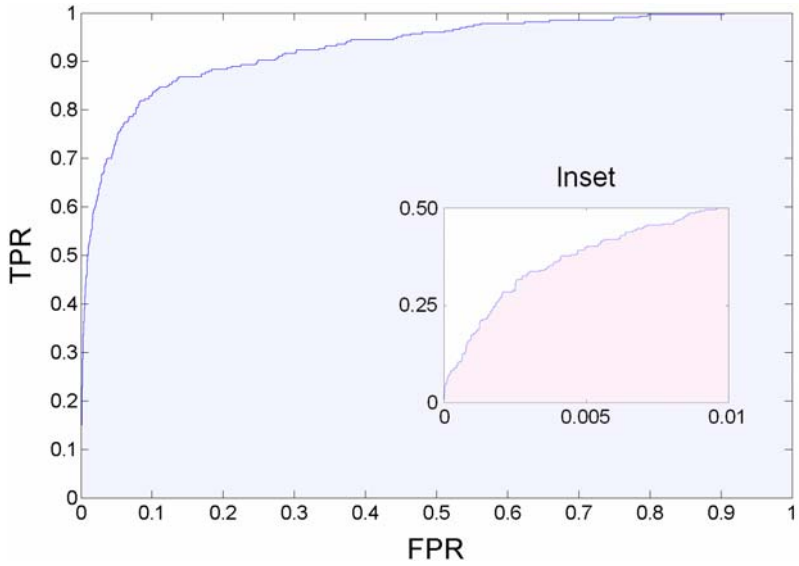


The affinity for the known *eve* regulators Bicoid (green line), Krueppel (blue line) and Hunchback (red line), respectively the combined affinity of these factors (black line) was calculated in 500 bp wide windows across the 18kb *eve* locus. The known stripes 1 through 7 and *ftz-like* are well detected by TRAP. Experimentally verified stripes as well as the coding region are indicated by black bars. Notice that affinity scores for different matrices live on similar scales.

5.2 P53 binding predictions

To assess how well the model can predict ChIP data from vertebrates I applied TRAP to the ChIP-PET data set of P53 from human cell-cultures (Wei et al., 2006). As explained in detail in Figure 2.13 the strength of TF binding to a given genomic region is indicated by the number of sequences that comprise a given PET cluster. The cluster size is thereby assumed to correlate about linearly with the binding strength of the TF to the corresponding genomic region. In addition, for the P53 data set the experimenters concluded that PET clusters of size ≥ 3 (PET3+) are highly indicative for binding of the TF to the corresponding genomic region while PET singletons are expected to represent almost exclusively random noise. Therefore all sequences spanning the 317 PET3+ clusters were used as true positives and 62.590 sequences corresponding to the PET singletons as true negatives (cluster sizes ranging in length from 37 to 6505 bases with an average of 635 bps). Alternatively, I selected 60.000 random genomic sequences with average length of the PET3+ clusters as a set of negatives but the results are minimally affected by this change. The quality of the TRAP

Figure 5.2 – ROC curve for the P53 ChIP-PET data set



The shape of the ROC curve with an AUC of 0.928 indicates high predictive power of TRAP over ChIP-PET data. The curve is based on 315 positive (PET3+) and 62.590 negative (PET singletons) sequences. The top 20 sequences according to ⟨N⟩ thereby contain 16 positives and only 4 negatives.

ranking for the PET clusters was subsequently evaluated using a ROC curve analysis analogous to the one outlined on page 81. As indicated by the large AUC of ~0.93 (Figure 5.2) TRAP allows identifying regions bound by P53 with high sensitivity and specificity. The results are further improved if only PET4+, PET5+ or PET6+ clusters are considered as the true positives suggesting that the predicted affinities are indicative of the relative binding strength. Assessing the Pearson correlation between PET size and ⟨N⟩ yielded a low but highly significant correlation of 0.184. This value is smaller than expected based on the very successful ranking. The reason for this lies in part in a group of PET clusters that are composed of centromeric and telomeric repeats. These clusters have no site with appreciable affinity for P53 but appear enriched after the antibody precipitation step only due to their high abundance in the genome. Since these PETs comprise the largest clusters the correlation coefficient is not particularly meaningful in this case.

5.3 Affinity ranking for the transcription factor SRF

The first part of this section focuses on the detection of target genes for the transcription factor SRF (Miano et al., 2007). In the second part the search is reversed and the question is addressed of whether TRAP can successfully identify SRF as the regulator of known SRF genes given a set of 593 different vertebrate PFMs from TRANSFAC (Matys et al., 2003).

In the previous sections the sequences to be used for computing the affinities of a given TF were nicely defined for instance by the 50 to 1500 bases long yeast intergenic regions spotted on the microarrays used in the chip studies. For the current analysis the situation is slightly more complicated as the genomic regions to be used for predicting the affinities of SRF for a given gene are rather undefined. Nevertheless, by relying on TSS annotations from experimental data (e.g. CAGE tags, Kawaji et al., 2006) or *ab initio* gene prediction programs (Uberbacher et al., 1991) one can define promoters likely enriched with sequence elements important for the regulation of the downstream gene. Following this logic,

Figure 5.3 – SRF gene locus

Gbrowse image of the genomic region from -10kb to +10kb around the TSS of the SRF gene. The upstream region was scanned by a 500 bp long window. From top to bottom, the displayed tracks correspond to GC content of the region, the gene structure of SRF, the sequence conservation between human and mouse and the CAGE tag support for the TSS (triangles), respectively. The bottom most track indicates the affinity of each 500 bp window. The window with highest affinity is located near the TSS verifying the presence of an experimentally known direct autoregulatory loop.

Table 5.1 – Top predicted SRF target genes

EnsEMBL ID	Gene Symbol	⟨N⟩	Distance
ENSG00000112658	**SRF**	**1.157**	**200**
ENSG00000184009	**ACTG1**	**1.133**	**74**
ENSG00000005981	ASB4	0.977	2155
ENSG00000184489	PTP4A3	0.949	870
ENSG00000120738	**EGR1**	**0.904**	**203**
ENSG00000075624	**ACTB**	**0.873**	**1274**
ENSG00000176895	OR51A7	0.872	8898
ENSG00000125848	FLRT3	0.859	5349
ENSG00000131437	**KIF3A**	**0.859**	**392**
ENSG00000117569	PTBP2	0.842	9013
ENSG00000117385	LEPRE1	0.836	3215
ENSG00000160808	**MYL3**	**0.829**	**10**
ENSG00000134460	IL2RA	0.821	65
ENSG00000141527	CARD14	0.819	8608
ENSG00000179388	**EGR3**	**0.808**	**164**
ENSG00000145425	RPS3A	0.788	8501
ENSG00000196413	ERVK6	0.787	3672
ENSG00000142871	CYR61	0.779	2259

Shown are the 15 genes with highest predicted affinity for the SRF matrix SRF_Q4. The first and second columns denote the Ensembl gene ID and the corresponding gene symbol. Column three indicates the maximal affinity found in the promoter. The last column shows the location of the window with largest affinity in respect to the TSS of the corresponding gene. Among the top genes according to affinity are many known SRF targets (indicated in bold). In contrast, when ranking according to the number of annotated hits as predicted by the balanced cutoff method the SRF gene is ranked only at position 816 together with 1044 other genes each having 5 annotated hits.

for each gene the sequence spanning the first 10 kb upstream of the corresponding TSS (according to Ensembl database, version 31) was defined as promoter region. Then, as outlined in Section 5.1, a 500 bp long window was shifted across the promoters and the affinity of each window was plotted against its start position. The result of this analysis is shown in Figure 5.3 for the gene encoding SRF itself. As indicated, the window with largest predicted affinity for SRF is located near the TSS verifying the presence of a well characterized direct auto-regulatory loop (Belaguli et al., 1997). Aside from pinpointing likely regulatory regions, the windows with maximal affinity were also used to rank all genes according to their affinity for SRF (see Table 5.1). Among the genes with highest predicted affinity is the gene encoding SRF itself as well as several other known SRF targets including actin and myosin encoding genes (Kumar et al., 1995, Zhang SX. et al., 2005) and the early growth response factors EGR1 and EGR3 (Shin et al., 2006, Tullai et al., 2004). As indicated by the location of the windows in respect to the TSS of the known SRF targets, nearly all high affinity sites with likely functionality are located within a few hundred base pairs upstream of the respective TSSs. The quality of the ranking could thus be further improved by narrowing down the search space from 10 kb to just 1 kb upstream of the TSS.

TF ranking for SRF target genes

In Section 4.2.2 it was shown that for a given intergenic region in yeast TRAP can oftentimes successfully predict the corresponding regulating TF by ranking all PFMs according to their predicted affinities for the region. Here I address the question of whether TRAP can also accurately predict the TFs regulating a given vertebrate gene or whether predicted affinities – obtained from combining the generic parameter description from yeast with vertebrate PFMs – reside on largely different scales.

To this end, for a given gene, its promoter was defined to comprise the genomic sequence spanning 1 kb upstream of the respective TSS and the affinities for all 593 PFMs from TRANSFAC were computed. The PFMs were subsequently ranked according to their predicted affinities for the gene. The results of this analysis are shown in Table 5.2 for the six genes encoding SRF, ACT1, EGR1, EGR3, CRX and E2F2. In accordance with experimental knowledge (Belaguli et al., 1997, Kumar et al., 1995, Shin et al., 2006, Tullai et al., 2004) and the previous findings the first four genes have SRF predicted as the top regulator (see Table 5.2). The latter two genes, which encode CRX and E2F2, served as a control as they are not SRF targets but rather encode other autoregulating transcription factors that bind directly to their own promoters (Nishida et al., 2003, Neuman et al., 1994). Also for these two genes TRAP predicted the corresponding PFMs among the matrices with highest affinity. In case of CRX this is particularly encouraging given the rather low information content of its corresponding PFM (6.4 bits).

Together these results indicate that predicted affinities for different vertebrate TFs are largely comparable when using the generic parameter description derived from yeast, irrespective of the length or information content of the corresponding matrices. This stays in stark contrast to the binding probabilities obtained from the simplified Boltzmann models, which, for different matrix lengths, reside on vastly different scales (data not shown). Also the ranking based on the number of annotated hits does in general not allow to detect the regulating TFs (see Table 5.2). Nevertheless, also the TRAP results may be further improved as the ranking of PFMs degrades noticeably when the promoter regions are further extended (data not shown). One solution to obtaining more robust and accurate rankings is provided by Manke, Roider and Vingron (2007) where we used Fourier transforms to derive the exact distribution of the binding affinities for a given TF and subsequently assigned p-values to any obtained affinity scores. These p-values allow to oftentimes accurately rank PFMs for a given gene even when extending the promoter region to several kilobases. Another implicit solution to the problem of ranking PFMs for a given sequence is provided by the methodology introduced in the next chapter.

Table 5.2 – TF ranking for individual 1kb promoters

	Gene					
	SRF	EGR1	ACT1	EGR3	CRX	E2F2
TRAP	TFIIA_Q6	SRF_Q6	SRF_C	SRF_Q6	WT1_Q6	AP2ALPHA_01
	SRF_Q4	SRF_Q5_02	SRF_Q5_01	SRF_C	OLF1_01	E2F_Q4
	SRF_Q5_01	SRF_Q5_01	SRF_Q5_02	SRF_Q4	CAP_01	E2F1_Q3
	SRF_Q5_02	SRF_Q4	SRF_Q4	E4F1_Q6	LRF_Q2	E2F_Q6
	WT1_Q6	SRF_C	AP2ALPHA_01	XPF1_Q6	UF1H3BETA_Q6	AP2_Q6_01
	HOXA3_01	WT1_Q6	SRF_Q6	SRF_Q5_01	TFIII_Q6	E2F_Q3_01
	OCT1_Q6	TFIII_Q6	SP1_01	SREBP1_Q6	STAT6_02	KAISO_01
	SRF_Q6	HES1_Q2	WT1_Q6	PAX4_03	HOXA3_01	E2F_02
	CACBINDING_Q6	AP2ALPHA_01	AP2_Q6	WT1_Q6	GEN_INI3_B	E2F1DP1RB_01
	RUSH1A_02	TFIIA_Q6	CHCH_01	ATF_01	CHCH_01	E2F1_Q4_01
	OCT1_Q5_01	LRF_Q2	LRF_Q2	CAP_01	HNF4_Q6_03	E2F1_Q6_01
	OCT_Q6	CHCH_01	ZF5_01	MUSCLE_INI_B	LEF1_Q2_01	E2F1_Q6
	STAT1_03	HOXA3_01	AP2GAMMA_01	RP58_01	GEN_INI2_B	MYOD_01
	LRF_Q2	ETF_Q6	HOXA3_01	HOXA3_01	SMAD_Q6_01	HOXA3_01
	CP2_01	SPZ1_01	E2F_Q2	NERF_Q2	CRX_Q4	E2F_Q4_01
Balanced cut-off	SP1_Q2_01	AP2_Q6_01	AP2_Q6_01	MUSCLE_INI_B	R_Q3	AP2ALPHA_01
	SP1_Q4_01	WT1_Q6	MUSCLE_INI_B	PAX4_03	MZF1_02	AP2_Q6_01
	SRF_Q4	MUSCLE_INI_B	AP2_Q6	PAX4_04	CAP_01	HIC1_03
	STAT6_02	SP1_Q4_01	SP1_Q6_01	SP1_Q4_01	ZIC3_01	AP2GAMMA_01
	GEN_INI_B	AP2_Q6	SP1_Q6	AP2_Q6_01	PAX2_01	OCT1_04
	SP1_Q6	SRF_Q5_02	SP1_Q4_01	SP1_Q6	MZF1_01	MUSCLE_INI_B
	TATA_01	SP1_Q6	SP1_01	STAT3_02	EGR_Q6	DR1_Q3
	SPZ1_01	SRF_Q4	E2F_Q2	ZIC2_01	NF1_Q6	DR4_Q2
	ZF5_01	SRF_Q6	ZIC3_01	GC_01	SP1_Q4_01	E2F1_Q6_01
	WT1_Q6	SRF_C	WT1_Q6	HIC1_03	SP1_Q2_01	ZF5_01
	STAT1_03	SP1_Q2_01	AP2GAMMA_01	HNF3_Q6	GATA4_Q3	R_Q6
	PAX4_03	MINI20_B	AP2ALPHA_01	YY1_01	WT1_Q6	STAT1_03
	MSX1_01	SRF_Q5_01	HIC1_02	SP1_Q6_01	TFIII_Q6	GEN_INI_B
	GEN_INI2_B	SPZ1_01	ZF5_B	SP1_Q2_01	LRF_Q2	GEN_INI3_B
	CAP_01	KROX_Q6	PAX4_03	SP1_01	SMAD_Q6_01	GEN_INI2_B

Shown are the 15 top ranking PFMs (out of 593 PFMs in TRANSFAC) for the indicated genes according to predicted affinity (top panel) or the number of annotated hits (bottom). In accordance with experimental findings, TRAP predicts SRF matrices as top ranking for the four SRF target genes ACT1, EGR1, EGR3 and SRF itself. For two other genes CRX and E2F2, which are not SRF targets but also possess a direct autoregulatory loop, the corresponding PFMs CRX_Q4 and several E2F matrices are found among the top 15 matrices while no SRF matrices are detected. In contrast, the balanced cutoff method detects fewer SRF matrixes for the SRF target genes, only one E2F matrix for the E2F promoter and the CRX matrix only at position 111 for the CRX gene.

Having demonstrated the applicability of the TRAP model to higher eukaryotes the next chapter will introduce two statistical methods that can be used to identify transcriptions factors that play an important role in the regulation of groups of genes. These methods will be used subsequently to perform a detailed large scale analysis of promoters from tissue specific genes.