CHAPTER 6

PASTAA – Predicting TFs associated with groups of genes

As shown in the previous two chapters, TRAP frequently detects known regulating factors for a given gene by ranking TFs according to their predicted affinities for the corresponding promoter. In this chapter I will extend this applicability of TRAP to the detection of TFs that play an important role in the regulation of groups of genes. The search for such TFs typically arises from expression studies (e.g. microarray or EST measurements) were changes in the transcriptome have been detected between different cell types or cell conditions and one seeks to understand the underlying regulatory processes responsible for the observed changes. Groups of genes with changed expression in a given condition are thereby assumed to share regulatory elements bound by the same transcription factor. While this appears to be a straight forward task, identifying the TFs responsible for establishing or maintaining the observed expression patterns constitutes a major challenge in bioinformatics.

Aside from the problem of correctly predicting the target genes of a TF, another difficulty is hereby the selection of meaningful gene groups to be analyzed. For instance, if we seek to find TFs that play an important role in the regulation of pancreatic genes we may want to select those genes which play a specific role in pancreas. However, in general there will only be a small number of genes whose expression is restricted to only a particular tissue and thus we need to introduce a cutoff that specifies how strongly the gene has to be upregulated in the tissue and in how many other tissues it may be expressed in order to declare it as sufficiently specific. Unfortunately, the thresholds yielding an optimal enrichment with direct targets for a given condition specific TF are not known *a priori* and may vary between different factors and conditions.

Another major obstacle for detecting regulating TFs in higher eukaryotes is the lack of well defined promoters. For instance, while the search for functional binding sites of a given TF to a yeast promoter can usually be limited to scanning the corresponding intergenic region of 50 to 1500 bps length, in vertebrates regulatory regions may lie kilobases away from the respective transcription start site. Solving this problem by simply annotating larger promoter regions is thereby impeded by the fast accumulation of random none functional binding sites, which blur any real binding signals.

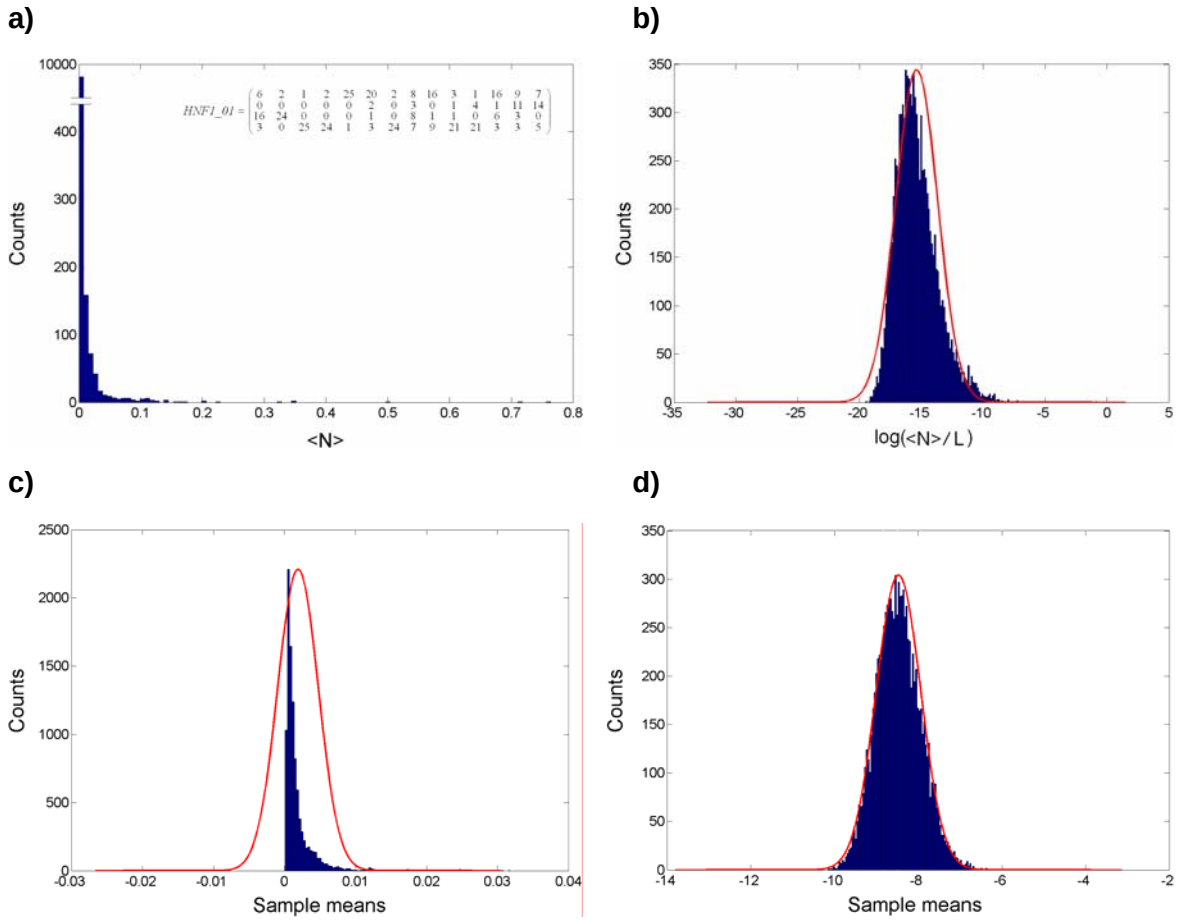
Despite the difficulties posed by selecting appropriate cutoffs for gene set construction, choosing suitable promoter regions and subsequently predicting functional binding sites, computational approaches led to the detection of a number of well-known regulatory associations between TFs and sets of target genes. Pioneering this work for tissue specific gene groups Wyeth Wasserman and colleagues (1998) were the first to detect the functional relationship between a small handpicked set of muscle and heart specific genes and TFs with known muscle specific function including SRF, MEF2 and MYOD. Subsequent large scale studies recovered several additional TF-tissue associations, such as the association between immune related factors and gene sets derived from leukocytes, and the association between hepatocyte nuclear factors and liver specific genes (Pennacchio et al., 2007; Smith et al., 2006; Yu et al., 2006). However, while each of these studies was able to recover at least a subset of the known associations, for many tissues and TFs no experimentally known associations have been recovered.

Here a new approach is introduce for detecting TF-gene set associations, which combines the affinity predictions of TRAP with two statistical tests that evaluate the enrichment of high affinity TF targets within a given input gene set. In case of non-categorical data such as presented by EST or microarray expression values, I introduce an iterative search for detecting optimal thresholds used to define the input gene sets. The performance of the resulting method, called PASTAA (**P**redicting **AS**sociated **T**ranscription factors from **A**nnotated **A**ffinities), will first be assessed on various gene sets with known common transcriptional regulators. It will then be used to perform a large scale promoter analysis of tissue specific genes where its performance will be compared to a number of state of the art methods.

6.1 Statistical measures for detecting affinity enrichments

In order to detect candidate transcription factors responsible for the regulation of a group of genes two statistical test schemes have been implemented. The first scheme is based on applying a z-score measure to detect increased affinity of a TF within the genes of an input set. The second scheme utilizes a series of hypergeometric tests designed to identify the enrichment of TF targets within the top ranking genes of a given input list. This test is specifically designed to deal with the fact that in most cases only a subgroup of the genes in the input list will actually be regulated by the same factor. In the following, the acronym PASTAA will refer to the combination of either of these statistical tests with the affinity predictions from TRAP.

Figure 6.1 – Affinity distribution for HNF1_01



a) For the 14 bases long matrix HNF1_01 raw affinities are log normal distributed over a set of 100.000 random sequences of length 1000bps. b) Log-transformed distributions for the affinities from a). c) Distribution of the sample means (sample size 30) when sampling form the original distribution. d) Sample means (sample size 10) approach a normal distribution (red lines) when sampling form the log-transformed affinity distribution. This allows deriving p-values directly from measuring the area under the Standard Normal Curve to the right of the z-score.

6.1.1 Testing for increased average affinity (z-score test)

The goal of detecting elevated average affinity of a given TF for the genes in a given input set pertains well to a z-test, which assesses whether a sample mean diverges significantly from the population average. In order to directly obtain valid p-values from the z-scores via the Standard Normal Distribution the sample means of the affinity values have to be normally distributed.

Affinities tend to be log-normally distributed

A typical histogram of the raw affinities over 10.000 randomly generated sequences (according to b = {0.25, 0.25, 0.25, 0.25} and length L = 1000 bps) is shown in Figure 6.1 for the matrix HNF1_01. As shown, the affinities are approximately log-normal distributed, that is, the distribution of the log-transfomred affinities approaches a normal curve. The sample means, computed from the transformed affinities for randomly generated gene sets, follow a Gaussian distribution even for small sample sizes (n = 10). In contrast, deriving sample means from raw affinites clearly violates the assumption of having a normal distribution even for sample sizes with n > 30. For details on the distribution of binding affinities I refer to Manke, Roider, Vingron (2007) where we analyzed the affinity distributions in detail, showing how they can be derived either analytically using characteristic functions or approximated using extreme value distributions. For the purpose at hand, the sample means of all factors approach a Gaussian distribution even for small sample sizes of $n \approx 10$ when using the log-transformation. z-scores obtained from such transformed affinities can thus directly be converted into valid p-values and be compared between factors, without the need to apply time-consuming resampling. On the other hand, transforming the affinities distorts their relative values, which might harm any subsequent analysis. Therefore in the following both, raw and transformed affinities were assessed for their applicability to the z-score test.

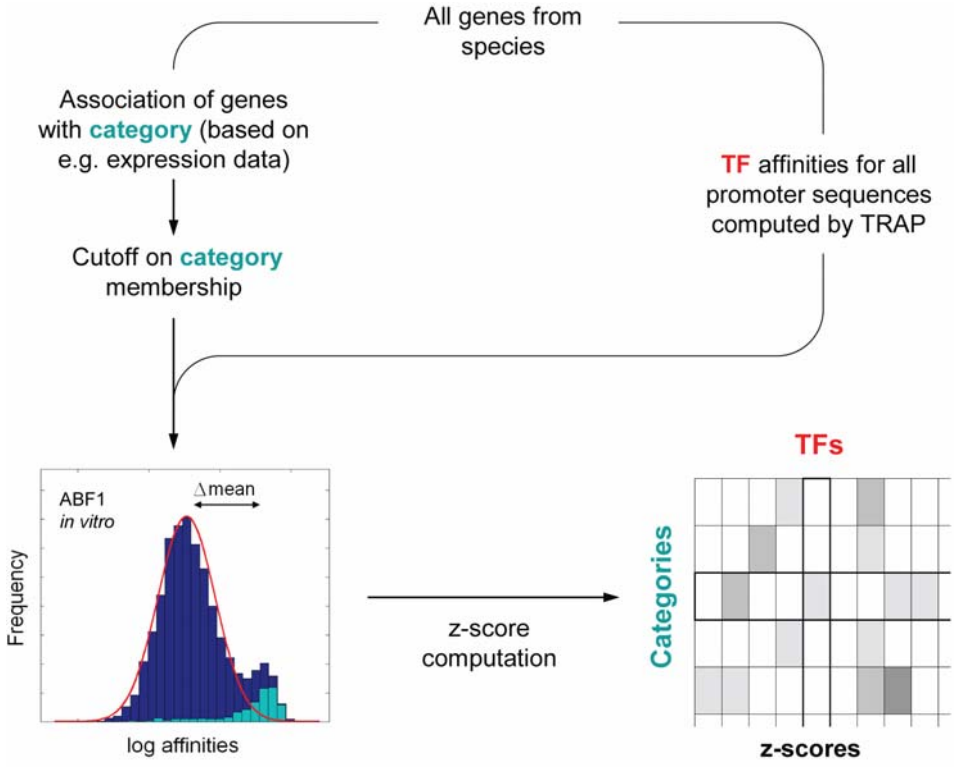
Applying the z-score test

Following the workflow outlined in Figure 6.2, once an input gene set has been chosen the following z-score (Rice, 1995) is computed using either the raw or log-transformed affinities from a given transcription factor:

$$z - score_{TF,Set} = \frac{\overline{\langle N \rangle}_{TF,Set} - \mu_{TF,Pop}}{\sqrt{\sigma^2_{TF,Pop} / n_{Set}}} \qquad (6.1)$$

where $\overline{\langle N \rangle}_{TF,Set}$ is the mean affinity of the TF for the promoters in the input gene set, $\mu_{TF,Pop}$ and $\sigma_{TF,Pop}$ are the mean and variance of the affinities over all promoters in the genome, and n_{Set} is the size of the input gene set. While using raw affinities allows to keep the relative binding affinities intact it makes the z-score test sensitive to outliers, that is, a single gene with exceptionally high affinity can cause a significantly elevated mean affinity in a given input set. Such outliers can occur for instance, when a PFM recognizes a repetitive sequence in a given promoter. Applying the log-transformation on the affinities diminishes the effect of outliers however, as mentioned before, also distorts the meaningful difference in predicted affinities between high and weak binding sites. In addition, the log-transform puts as high weights on low affinities as it does on high ones (both tails of the Gaussian

Figure 6.2 – Workflow for PASTAA using the z-score test



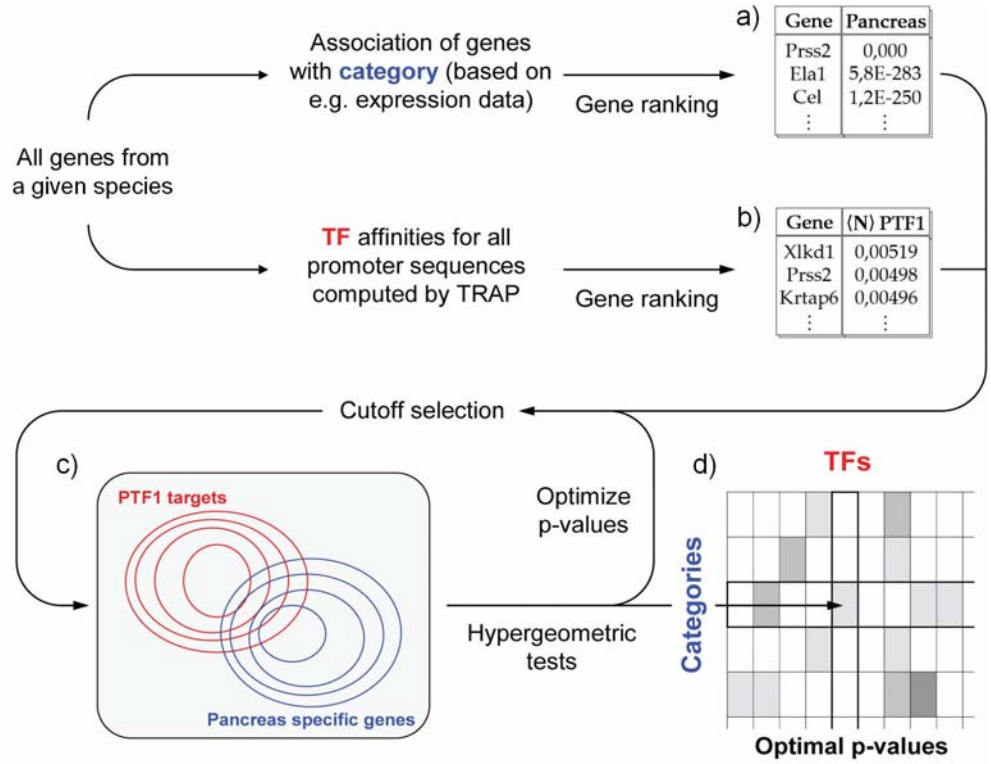
All genes of a given species are separated into categories (input gene sets) based on for instance EST expression data or ChIP-chip binding p-values (indicated for the Abf1 in vitro binding data). For a given TF the mean affinity of the promoters in each such category is then compared to the mean affinity of the population of all promoters in the genome. The computed z-scores are subsequently used to rank all TFs for a given category or alternatively, all categories for a given TF.

distribution contribute equally to the mean). In turn the average for a given input set might not be significantly elevated due to the presence of genes with spuriously low affinity. To avoid this dilemma and to deal with the fact that input gene sets are oftentimes badly defined a method relying on a series of hypergeometric tests, as described below, was implemented.

6.1.2 Testing for target gene enrichment (hypergeometric test)

As mentioned above, spuriously low or high affinity predictions for individual genes can negatively influence the results of the z-score test. To avoid this problem I alternatively test for the enrichment of TF targets among the genes of an input set. To this end, for a given PFM all genes are first ranked according to their predicted affinity. Subsequently, a cutoff is applied to the ranked list that separates all genes in the genome into targets and non-targets

Figure 6.3 – Workflow for PASTAA using the hypergeometric test



All genes of a given species are ranked according to their association with a given category such as pancreas a). At the same time the genes are also ranked according to the predicted affinity for a given TF such as PTF1 b). After applying a cutoff to the lists in a) and b) a hypergeometric test is used to determine the overlap between the top ranking genes of the TF and the top ranking genes in the category (illustrated by the Venn diagram in c). Cutoffs are thereby chosen iteratively in such a way that the obtained hypergeometric p-values (ovals indicate the corresponding changes in set sizes) are minimized. The significance of the associations between all PFMs and the provided categories is stored in a matrix as depicted in d) where grey scales indicate the level of significance of a given TF-category association.

of the corresponding TF. Finally, following the scheme shown in Figure 6.3, a hyper-geometric test is employed:

$$P(x \geq X) = 1 - \sum_{k=0}^{X-1} \frac{\binom{T}{k}\binom{\mathrm{N}-\mathrm{T}}{\mathrm{C}-\mathrm{k}}}{\binom{\mathrm{N}}{\mathrm{C}}} \qquad (6.2)$$

which computes the probability of observing exactly *k* or more genes in the intersection between the input set of size *C* and the target set of size *T*, given a total number of *N* genes.

Cutoff selection for TF targets and tissue specific genes

The significance of the association between a given TF and the genes in the input set, as obtained from the hypergeometric test, depends on the cutoff used to select the genes for the input set and on the cutoff on ⟨N⟩ used to specify the targets of a given TF. Since the optimal values for these two thresholds are not known *a priori* PASTAA loops over a set of cutoffs on both, the values that determine association with an input set (e.g. significance of the expression of a gene in a tissue), as well as on ⟨N⟩. For the input set the cutoff is chosen in such a way that sets containing {1, 2, … , 99, 100, 110, 120, … , 290, 300, 400, … , 900, 1000} genes are generated. On ⟨N⟩ the threshold is chosen so that target gene sets of size {25, 50, 75, 100, 125, 150, 175, 200, 250, 300, 350, 400, 450, 500, 600, 700, 800, 900, 1000} are obtained. The choice to put the maximal set size to 1000 genes will be explained in context of Section 6.3.1. In general, each of the resulting 2413 cutoff combinations will yield a different hypergeometric p-value. I assume that the smallest achieved hypergeometric p-value thereby corresponds to the most meaningful detectable association between a given TF and a set of genes. The obtained p-values are directly used to rank the TFs for a given input list. However, due to the apparent multiple testing problem the p-values need to be corrected if one seeks to accurately determine the significance of a given association. This is done by applying 100.000 rounds of resampling for any given input set size. For each resampling the gene labels are thereby shuffled randomly between all *N* genes of the genome before performing the analysis.

Next to expression data, groups of genes could also be derived from categorical data as presented by databases such as KEGG (Aoki et al., 2005) or Gene Ontology (Hill et al., 2002). In such a case one might seek to find TFs that regulate the expression of genes unambiguously assigned to a particular metabolic pathway such as glycolysis or a process such as DNA repair. The genes belonging to the corresponding categories are in general not ranked and are thus all treated equal, that is, no additional cutoff is applied to the input set.

It has to be stressed that the cutoffs on ⟨N⟩ are introduced not on the level of individual binding sites but on the affinity scores for entire promoters. While this still constitutes an arbitrary separation between targets and non-targets, the relative contributions of the binding sites to ⟨N⟩ are retained, which allows to take advantage of the improved gene ranking introduced by TRAP. In principle, the method of varying cutoffs can also be applied to define the input gene set for the z-score test. However, due to the problem of outliers in the affinity of individual promoters, the minimal input set size has to be chosen rather large in order to avoid spuriously high z-scores for small input sets. This in turn limits the usability of

flexible thresholds. This is underlined by the fact that for the following analyses I was not able to significantly improve the TF rankings obtained from the z-score test by applying flexible cutoffs (data not shown).

6.2 Data and methods used for analyzing tissue specific gene sets

Section 6.5 deals with a detailed analysis of tissue specific genes and TFs. This analysis requires, next to the definition of tissue specific input gene sets, also the generation of sets of vertebrate promoter sequences. In the following it will briefly be outlined how the sequences and expression data used for the tissue analysis were obtained and how to detect a general expression bias of TFs across different tissue categories.

6.2.1 Sequence data

All mouse and human genomic sequences as well as the annotation of the corresponding transcription start sites (TSSs) for 26.609 mouse and 30.423 human genes were taken from the Ensembl database version 31 (Birney et al., 2006). As representation of the promoter regions of the genes two different sequence sets were generated. The first set consists of the full genomic mouse sequence up to 10kb upstream of each TSS. The second set consists of all sequence blocks conserved between the mouse and human genes with annotated one to one orthology relationship in Ensembl. The alignments, as provided by Stefan Haas, were generated from a region of 20kb upstream of the respective TSSs and applying the BLASTZ algorithm (Schwartz et al., 2003). Repetitive sequences were thereby masked beforehand using RepeatMasker (Smit et al., 1996).

6.2.2 Defining tissue specific gene expression

EST data

The expression of a given gene in a given tissue from human and mouse was determined by analysing corresponding EST clusters from the database GeneNest (see Section 2.4.2, Haas et al., 2000), which includes the annotation of the originating tissue for each EST. To detect EST clusters whose distribution of ESTs derived from various tissues differ significantly from the expected distribution all clusters were subjected to a binomial test such that a p-value describing the likelihood of observing a given number of ESTs from a given tissue in an EST cluster of given size was obtained. These EST cluster p-values reflect the extent of

overexpression of a given gene in a given tissue and were successfully used previously to predict tissue-specific expression of genes (Gupta et al., 2005). To minimize the problem of insufficient EST sampling for genes the tissue lists from human and mouse were combined by always taking the more significant p-value for each orthologous gene and tissue from either species. To speed up computations and to allow for a meaningful comparison with alternative approaches that do not adjust the input set size (PAP, Clover, oPOSSUM) only EST clusters with p-value < 10^{-6} in at least one of the 72 tissue categories were used subsequently as input to the statistical tests. For the hypergeometric test all genes in a given set were ranked according to their tissue p-values.

GNF data

Alternatively, I retrieved the expression values from the mouse GNF data set (Su et al., 2004) that contains MAS5 treated expression values for 12.191 Ensembl mouse genes in 54 tissue categories, 32 of which have a corresponding category in the EST data set. The expression values were normalized by a z-score transformation:

$$z_{t,g} = \frac{\Phi_{t,g} - \overline{\Phi}}{\sigma} \qquad (6.3)$$

where $z_{t,g}$ is the normalized expression level, $\Phi_{t,g}$ is the expression value of gene g in tissue t, $\overline{\Phi}$ is the average expression over all tissues and σ is the variance of the expression of gene j over all tissues. In the following, a given tissue input gene set for PASTAA consisted of all genes with a $z_{t,g}$ value > 0 for the corresponding tissue. For the hypergeometric test all genes in a given input set were ranked according to their $z_{t,g}$ values for the tissue.

6.2.3 Assessing TF expression across tissues

To test whether TFs are in general preferentially expressed in the tissues most significantly enriched with their target genes I first select for each TF-tissue association the PFM, which yields the most significant hypergeometric p-value with any of the tissues. This is done to avoid any bias possibly introduced by having several PFMs for a given TF. Then to determine the expression level of a given TF in a given tissue the protein sequence of the TF, as provided by TRANSFAC, was mapped to the mouse or human EST cluster with highest sequence similarity according to BLASTX (mapping provided by Stefan Haas). TFs with EST cluster p-value < 10^{-6} in the corresponding tissue were selected as specifically expressed. Subsequently, all cases where a TF is specifically expressed in its top ranking tissue were put in a first bin, all cases where a TF is specifically expressed in its second to top tissue in a second bin and so forth. For each TF this procedure was repeated over all its

72 tissue associations. The ultimate assessment of the size of the resulting bins is complicated by the fact that tissue categories with few ESTs are not only less likely to express the TF but are also less likely to produce significant hypergeometric p-values. Therefore, there exists an intrinsic negative correlation between the ranks of the tissue and the number of TFs expressed per tissue. To assess whether the enrichment is higher than expected by chance, I repeated the entire analysis ten times, every time assigning a random 200 bp long DNA sequences to each of the 26.609 Ensembl gene IDs (alternatively one may use a resampling procedure that randomly shuffles the gene identifiers between all genes in the genome). The difference between the actual results and the ones obtained from the random sequences in each of the 72 bins was finally evaluated by computing the following t-statistic:

$$t_i = \frac{bin_{i,g} - \overline{bin}_{i,r}}{\sigma_r} \qquad (6.4)$$

where $bin_{i,g}$ is the number of TFs assigned to bin *i* using the real genomic sequences, $bin_{i,r}$ is the average number of TFs assigned to bin *i* over all ten random sequence sets and σ_r is the standard deviation of the number of TFs in bin *i* obtained over the 10 random sets.

6.3 Validation of the PASTAA approach

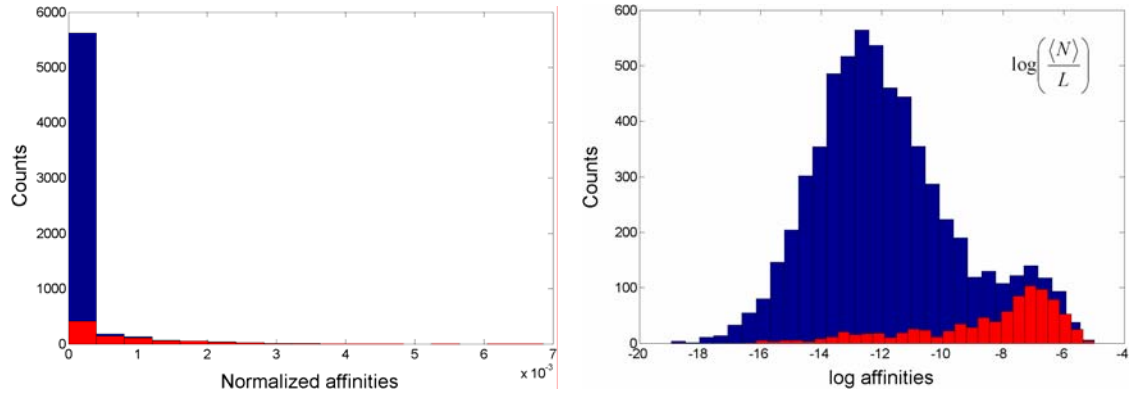
To assess the validity of the outlined statistical measures the z-score and hypergeometric tests is first applied to the same PBM and ChIP-chip data sets from yeast already used in Chapter 4 (for details on the data sets see Section 4.1.3). The performance of the tests is compared to the results of a simple ROC curve analysis. Subsequently, the validation of the hypergeometric test is extended to vertebrate data sets with known associated TFs. A comprehensive comparison with alternative state of the art approaches will be provided in Section 6.4.5.

6.3.1 Performance on ChIP-chip and PBM data sets from yeast

The ChIP-chip and PBM data sets from yeast constitute an ideal test case for evaluating PASTAA's sensitivity and specificity when searching for elevated average affinity of a given TF or an enrichment of predicted TF targets in an input gene group.

Figure 6.4 – Affinity distribution for ABF1_01

a) **b)**



a) Raw affinities distribution from ABF1_01 over all 6725 intergenic regions from yeast, normalized by the length of the corresponding region. b) log-transformed affinity distribution. Red highlights the intergenic regions significantly bound in the PBM experiment.

Results obtained from the z-score test

When using the z-score statistics PASTAA requires a precise definition of the input gene set. To this end, all genes with ChIP-chip binding p-value < 10^{-2} were defined as the true targets of a given TF. As an illustrative example, Figure 6.4 shows the distribution of the raw and log-transformed affinities for the matrix ABF1_01, after normalizing by the length of the intergenic regions. As indicated, the 885 sequences bound by ABF1 according to *in vitro* PBM binding data have a mean affinity strongly shifted towards higher values of ⟨N⟩. In this case, the average raw affinity of the population of all 6725 genes is 7×10^{-5} while the average raw affinity of the set of significantly bound sequences is 3.1×10^{-4}. With a population variance of 2.9×10^{-4} a z-score of 21.4 was obtained. To convert this z-score into a valid p-value I applied 100.000 rounds of resampling, none of which resulted in a larger z-score. The significance of this z-score is thus < 10^{-5}. Alternatively, using the log-transformed affinities yields a z-score of 36.8, which directly corresponds to a highly significant p-value of 5.5×10^{-281}. Both procedures indicate that the mean affinity of the input set is very much different from the average over all genes. While it is encouraging to see that ABF1_01 obtains a significant p-value for the ABF1 chip data set it remains unclear if this matrix exhibits the strongest association with the data set or whether other matrices yield even more significant p-values. To assess this issue I used either the raw or the log-transformed affinities to obtain the z-scores for the association of all 110 fungi PFMs in TRANSFAC with the PBM gene set of ABF1. By ranking all PFMs according to their z-scores derived from raw affinities the matrix ABF_C was found to be top ranking (z-score 29.7) followed by TAF_Q6, another matrix

Table 6.1 – Top factors for a given chip data set according to the z-score test

CONDITION	RANK				
	1	2	3	4	5
ABF1 *in vitro*	ABF_C	ABF1_01	N/A	N/A	N/A
ADR1 YPD	30	N/A	N/A	N/A	N/A
CBF1 SM	CBF1_B	N/A	N/A	N/A	N/A
GAL4 galactose	GAL4_01	GAL4_C	LAC9_C	N/A	N/A
GCN4 SM	GCN4_01	GCN4_C	LEU3_B	N/A	N/A
GCR1 YPD	GCR1_B	GCR1_01	N/A	N/A	N/A
HAC1 YPD	HAC1_Q2	N/A	N/A	N/A	N/A
HAP1 YPD	HAP1_B	N/A	N/A	HAP234_01	N/A
HAP4 rapamycin	HAP234_01	N/A	N/A	N/A	N/A
HSF1 H_2O_2Lo	HSF_04	N/A	HSF_01	N/A	N/A
LEU3 SM	LEU3_B	N/A	GCN4_01	N/A	GCN4_C
MATA1 YPD	31	N/A	N/A	N/A	N/A
MCM1 α Factor	MCM1_01	MCM1_02	N/A	N/A	N/A
MIG1 *in vitro*	MIG1_01	N/A	N/A	N/A	N/A
MSN2 H_2O_2Hi	HSF_04	STRE_B	N/A	STRE_01	N/A
PDR3 YPD	N/A	PDR3_Q2	N/A	N/A	N/A
PHO4 phosphate	PHO4_01	N/A	N/A	N/A	N/A
ROX1 H_2O_2Hi	30	N/A	N/A	N/A	N/A
RAP1 *in vitro*	RAP1_C	N/A	N/A	N/A	N/A
RCS1 H_2O_2Hi	RCS1_Q2	N/A	N/A	N/A	N/A
REB1 YPD	REB1_B	N/A	N/A	N/A	N/A
STE12 α Factor	STE12_Q4	MCM1_01	N/A	STUAP_01	MATα2_01
XBP1 H_2O_2Lo	N/A	N/A	XBP1_Q2	N/A	N/A
ZAP1 YPD	ZAP1_Q6	N/A	N/A	N/A	N/A

The first column indicates the ChIP-chip or PBM data set from which the gene set was derived. All genes with ChIP p-value < 10^{-2} were thereby assigned to the respective sets. Columns 2 to 6 show the top ranking PFMs for the given condition. Red and yellow indicate matching matrices and likely co-regulators, respectively. For instance, both LEU3 and GCN4 are involved in controlling amino acid synthesis, HSF and MSN2 cooperatively regulate the expression of various stress response genes and MCM1 and STE12 are directly interacting factors involved in pheromone signaling. Not further assessed associations are denoted as N/A.

describing the ABF binding motif (z-score 27.3), and third ABF1_01. In contrast, the next matrix, REPCAR1_01, had a z-score of 4.2 demonstrating a strong separation between the ABF matrices and all others. The same ranking was obtained when using z-scores from log-transformed affinities.

To test the overall performance of the method the analysis was extended to all the PBM and ChIP-chip data sets from Chapter 4. The ranking of individual PFMs for these data sets is shown in Table 6.1 for the z-score test applied to raw affinities. In 18 out of 24 cases the correct PFM is identified as the most significantly associated matrix for the corresponding chip data set. In addition to these perfect matches, the z-score test detects several known co-factors for many of the TFs. For instance, LAC9, a known co-regulator for galactose response genes (Salmeron et al., 1989), is the second highest ranked matrix for the GAL4 data set after GAL4 itself. For the data set from heat shock factor (HSF) a matrix corresponding to MSN2/4 is identified, a known co-regulator of HSF controlled genes (Grably et al., 2002). Another example is MCM1 and STE12, two directly interacting factors involved in the response to pheromone signaling (Primig et al., 1991). Accordingly, the PFMs for STE12 and MCM1 are found as top matrices for the STE12 data set. Taken together, meaningful associations are found for all but three factors (88% of cases). Importantly, this also includes factors such as PHO4 and PDR3 for which in Chapter 4 no significant correlation could be detected between the TRAP predictions and the experimental *R/G* ratios due to noise in the ChIP data. The only ChIP-chip data sets where no corresponding matrices could be recovered are ADR1 and MATα1 in YPD condition and ROX1 in oxidative stress condition. ADR1 is involved in response to glucose deprivation and thus likely not bound to its target genes in rich medium (Simon et al., 1991). MATα1 represses mating type genes in a/α diploid cells genes but is quickly degraded in haploid cells (Johnson et al., 1998). Since MATa cells were used by Harbison et al. (2004, http://jura.wi.mit.edu/young_public/regulatory_code/OSStrainList.xls) the protein was likely not present at the promtoers of its target genes. ROX1 however, is expected to be bound to its target genes under oxidative stress and thus constitutes the only false negative prediction. The z-score test based on raw affinities thus performs with high specificity and sensitivity, which suggests that such z-scores from different factors are well comparable to each other. In addition, converting raw affinity z-scores into p-values via resampling did not further improve the rankings in any apparent way (possibly due to the maximal p-value resolution that can be obtained by 10^5 resamplings). Interestingly, using z-scores based on log-transformed affinities recovered significantly less of the known TF-chip-set associations indicating that it is more important to keep the relative affinity values intact rather than to normalize the affinity distributions.

Having established that z-scores from raw affinities are well suited for ranking PFMs for a given gene set I now ask the reverse question, that is, given a PFM, can these z-scores be used to identify the corresponding chip data sets? In real world applications this corresponds to the case where a TF matrix is known and one wants to find the biological

Table 6.2 – Top ranking ChIP-chip set for a given TF according to the z-score

MATRIX	RANK				
	1	2	3	4	5
ABF1_01	YPD	*in vitro*	N/A	N/A	N/A
ABF_C	YPD	*in vitro*	N/A	N/A	N/A
ADR1_01	N/A	N/A	N/A	N/A	N/A
CBF1_B	YPD	SM	N/A	N/A	N/A
GAL4_01	YPD	raffinose	galactose	MIG1 gal.	N/A
GAL4_C	YPD	raffinose	galactose	MIG1 gal.	N/A
GCN4_01	rapamycin	SM	YPD	N/A	N/A
GCN4_C	SM	rapamycin	YPD	LEU3 SM	LEU3 YPD
GCR1_01	YPD	N/A	N/A	N/A	N/A
GCR1_B	YPD	N/A	N/A	N/A	N/A
HAC1_Q2	YPD	N/A	N/A	N/A	N/A
HAP1_B	YPD	HAP4 YPD	N/A	N/A	N/A
HAP234_01	HAP4 YPD	HAP4 rapa.	HAP2 YPD	HAP3 YPD	HAP1 YPD
HSF_04	H_2O_2Lo	H_2O_2Hi	MSN4 H_2O_2Hi	MSN2 H_2O_2Lo	MSN2 acid
LEU3_B	SM	YPD	GCN4 SM	N/A	N/A
MATA1_01	N/A	N/A	N/A	N/A	N/A
MCM1_01	α Factor	YPD	STE12 but. 14	STE12 but. 90	N/A
MCM1_02	α Factor	YPD	STE12 but. 14	STE12 but. 90	N/A
MIG1_01	*in vitro*	galactose	GAL4 raff.	N/A	N/A
STRE_B	MSN2 H_2O_2Hi	MSN4 H_2O_2Hi	MSN2 H_2O_2Lo	MSN2 acid	N/A
STRE_01	MSN4 H_2O_2Hi	MSN2 H_2O_2Hi	MSN2 H_2O_2Lo	MSN2 acid	N/A
PDR3_Q2	N/A	N/A	N/A	N/A	N/A
PHO4_01	phosphate	N/A	N/A	N/A	N/A
ROX1_Q6	N/A	N/A	N/A	N/A	N/A
RAP1_C	*in vitro*	FHL1 SM	FHL1 rapa.	FHL1 YPD	SM
RCS1_Q2	H_2O_2Hi	N/A	N/A	N/A	N/A
REB1_B	H_2O_2Lo	YPD	H_2O_2Hi	N/A	N/A
STE12_Q4	α Factor	butanol 90	YPD	butanol 14	MCM1 YPD
XBP1_Q2	H_2O_2Lo	N/A	N/A	N/A	N/A
ZAP1_Q6	YPD	N/A	N/A	N/A	N/A

Columns 2 to 6 denote which ChIP data set had largest z-scores for the PFM indicated in column 1. Red highlights matching TF-data set pairs; yellow indicates sets that correspond to known co-regulating TFs. For example, both RAP1 and FHL1 are involved in ribosomal gene regulation and share many targets. Accordingly, PASTAA identifies the RAP1 and FHL1 ChIP data sets as most significantly enriched with predicted RAP1 target genes. Not further assessed associations are denoted as N/A.

context in which the factor most likely plays a role. To assess the quality of this reverse search z-scores are used to rank all ChIP-chip and PBM data sets for a given PFM. The results of this ranking are shown in Table 6.2 for the 30 TRANSFAC PFMs for which chip data is available. The ranking of the test data sets for a given PFM again proofs to be successful. For 26 out of the 30 matrices the corresponding chip data sets are identified as top associated. Interestingly, the ranking often resembles what is known about the activity of the TFs. For instance, for HSF (heat shock factor) and MSN2/4 (STRE), two factors crucial also for the response to oxidative stress (Hashikawa et al., 2006), the appropriate stress conditions are top associated while for LEU3 and GCN4, two factors involved in controlling amino acid synthesis (Wang et al., 1999), the data set from amino acid starvation is top ranking. In addition, for many PFMs the chip data sets from known co-factors are among the top ranking associations. Somewhat unexpectedly, for many of the stress response factors also the corresponding YPD chip data sets are identified. This indicates that many of the stress response factors are bound to their target genes also in rich medium condition (YPD) although perhaps without the presence of a required cofactor.

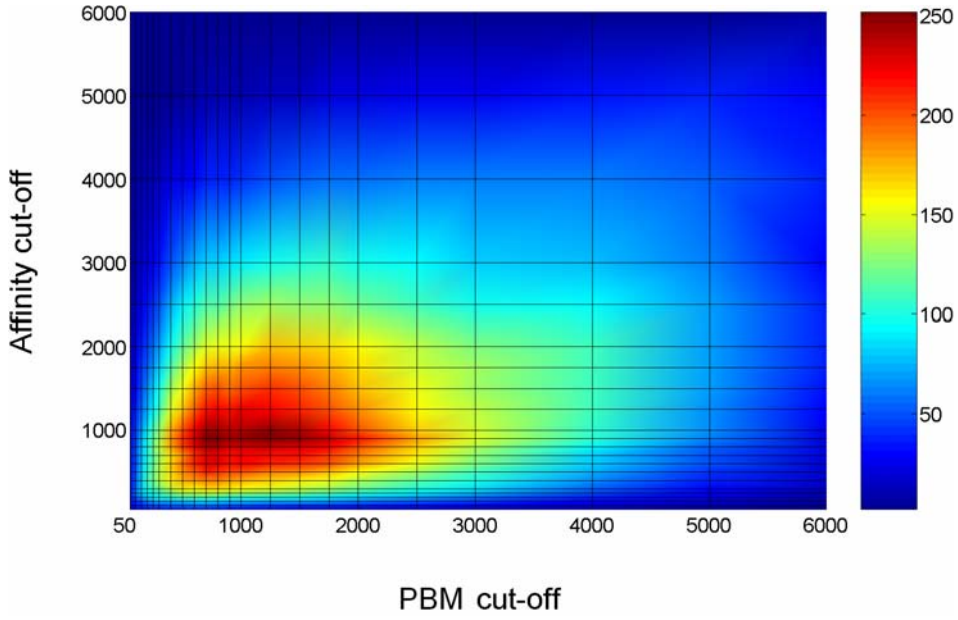
Results obtained from the hypergeometric test

When using the hypergeometric test statistics PASTAA adjusts the size of the input gene set as well as the number of genes predicted as targets for a given TF automatically in such a way that the enrichment of predicted targets among the input genes is optimized. Therefore, the explicit definition of a cutoff on PBM and ChIP-chip binding p-values is not applied but instead all genes, ranked according to chip binding p-values, are supplied to PASTAA.

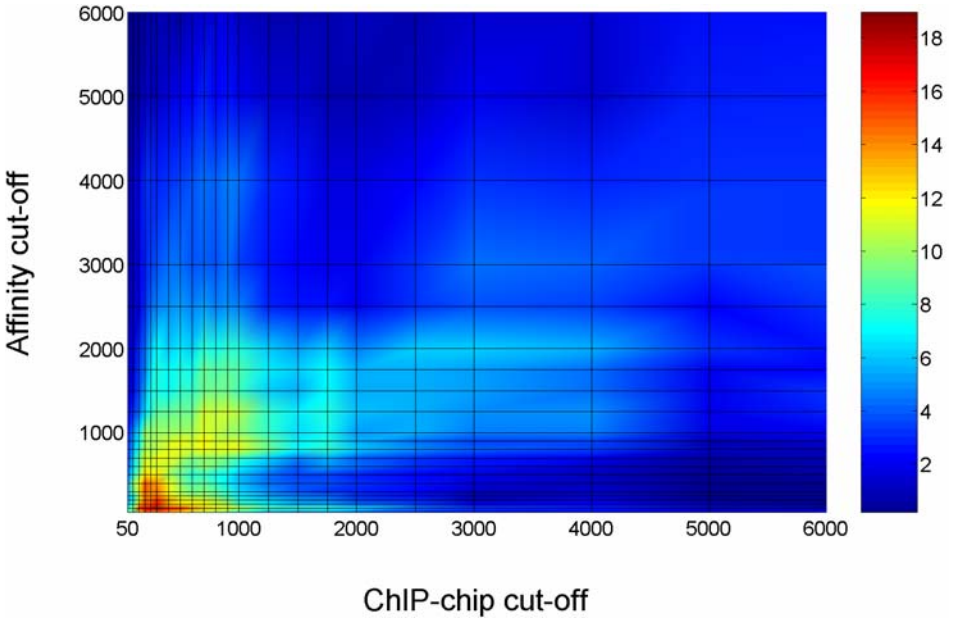
When performing the analysis for ABF1_01 and the corresponding PBM data set a hypergeometric p-value of 7.3×10^{-253} is obtained (ignoring multiple testing) for the enrichment of predicted TF targets among the genes of the optimized input set. This enrichment is found when using an optimized input gene set corresponding to the top 800 intergenic region according to chip binding p-values and a target set consisting of the top 900 genes according to $\langle N \rangle$. The resulting sets share a total of 474 intergenic regions. The behaviour of the hypergeometric p-values across the entire cutoff space is shown in Figure 6.5a for ABF1_01 and the corresponding PBM data set. Most significant p-values are found around the line of 900 target genes and between 700 and 1500 input genes. Importantly, taking the derivative of the cutoff space shows that the hypergeometric p-values grow fastest near the origin of the plot, indicating that the genes with highest affinity are in fact the most likely ones to belong to the input set (data not shown). In addition, the cutoff space reveals a significant association between predicted and real targets even if about half of the intergenic regions of the yeast genome are considered as input and target sets, respectively, indicating a high

Figure 6.5 – Cutoff space for the hypergeometric test

a)



b)



a) Shown are the -log hypergeometric p-values for ABF1_01 and the ABF1 *in vitro* data set depending on the cutoff combination employed for the predicted affinity and PBM binding values. The most significant target enrichment (p-value 7.3×10^{-253}) is found when using the top 800 genes according to PBM and top 900 genes according to affinity. The steepest increase in -log p-values is found at the origin of the plot. b) Same analysis as in a) but for the factor PHO4_01 and the PHO4 ChIP-chip data set. According to the fact that PHO4 has far less targets than ABF1 an optimal hypergeometric p-value = 7.9×10^{-20} is found when using only the top 300 genes according to ChIP-chip data and top 100 genes according to affinity.

robustness of the hypergeometric test against including false positives in either gene set. Given that most factors have considerably less real targets than ABF1, which is a global transcriptional regulator involved in the regulation of a multitude of genes (Miyake et al., 2004), the optimal cutoffs for more specifically acting factors are expected to lie below 1000 genes for both the target and input set. In fact, all other tested yeast matrices (30 from TRANSFAC and 4 matrices derived by the Fraenkel Lab, Harbison et al., 2004) had optimal cutoffs below 1000 genes. This is illustrated exemplarily in Figure 6.5b for the matrix PHO4_01 and its corresponding ChIP-chip data set. For this factor, with only around 40 target genes (Gonze et al., 2005), the optimal p-value of 7.9×10^{-20} is found when using the top 300 genes according to ChIP-chip data and top 100 genes according to affinity.

The time complexity for an individual hypergeometric test grows with *T,* the number of target genes, and *k*, the number of genes shared between input set and target gene set according to $O(k(T+k))$. In the following the maximal set sizes are thus restricted to 1000 genes, which greatly speeds up the computation times. The efficiency of the computation can further be improved by limiting the input sets *a priori* to only the significantly bound genes. In this case the cutoff procedure applied by PASTAA merely refines the predefined input set to obtain again optimal p-values. For instance, when given the same PBM input genes as supplied to the z-score test (885 intergenic regions bound by ABF1 *in vitro*) then this set is refined to contain 700 genes while the target gene set is adjusted to contain 600 genes. This yields an overlap of 358 genes between the two sets, which corresponds to a hypergeometric p-value of 6.8×10^{-229}. It should be noted that supplying *a priori* information regarding the input gene set also minimizes the risk of running into spurious p-value minima somewhere across the cutoff space. When applicable, such *a priori* knowledge is thus incorporated in the analyses presented in subsequent section.

When applying the hypergeometric test procedure to all the PBM and ChIP-chip data sets from Chapter 4 PASTAA finds in 19 out of 24 cases the correct PFM as the most significantly associated matrix for the corresponding chip data set. In addition, as for the z-score test, for many transcription factors known co-factors are recovered. For instance, MIG1 is identified as 4th matrix for the GAL4 data set after GAL4_01, GAL4_C and LAC9 (Nehlin et al., 1991). The complete list of top ranking associations is shown in Table 6.3. Overall, the outcome closely resembles what was found by the z-score test. Interestingly, also the results of the reverse search (ranking the data sets for a given PFM) closely resemble what was obtained by the z-score test (see Table 6.4). This is in line with the expectation that the hypergeometric p-values obtained for different PFMs can directly be compared. Together these findings suggest that both statistical methods are well suited for detecting TF-gene set

Table 6.3 – Top factors for a given chip data set according to hypergeometric tests

CONDITION	RANK				
	1	2	3	4	5
ABF *in vitro*	ABF1_01	ABF_C	N/A	N/A	N/A
ADR1 YPD	59	N/A	N/A	N/A	N/A
CBF1 SM	CBF1_B	N/A	N/A	N/A	N/A
GAL4 galactose	GAL4_C	GAL4_01	LAC9_C	MIG1_01	N/A
GCN4 SM	GCN4_01	GCN4_C	LEU3_B	N/A	N/A
GCR1 YPD	GCR1_01	GCR_B	N/A	N/A	N/A
HAC1 YPD	8	N/A	N/A	N/A	N/A
HAP1 YPD	HAP1_B	N/A	N/A	N/A	HAP234_01
HAP4 rapamycin	HAP234_01	N/A	N/A	N/A	N/A
HSF H_2O_2Lo	HSF_04	HSF_05	HSF_03	STRE_B	STRE_01
LEU3 SM	LEU3_B	GCN4_01	GAL4_C	GCN4_C	N/A
MATA1 YPD	14	N/A	N/A	N/A	N/A
MCM1 α Factor	MCM1_01	MCM1_02	N/A	N/A	N/A
MIG1 *in vitro*	MIG1_01	N/A	N/A	N/A	N/A
MSN2 H_2O_2Hi	STRE_B	STRE_01	ADR1_01	HSF_04	N/A
PDR3 YPD	N/A	N/A	N/A	PDR3_Q2	N/A
PHO4 phosphate	PHO4_01	N/A	N/A	N/A	N/A
ROX1 H_2O_2Hi	90	N/A	N/A	N/A	N/A
RAP1 *in vitro*	RAP1_C	N/A	N/A	N/A	N/A
RCS1 H_2O_2Hi	RCS1_Q2	N/A	N/A	N/A	N/A
REB1 YPD	REB1_B	N/A	N/A	N/A	N/A
STE12 α Factor	STE12_Q4	MCM1_01	MATα2_01	N/A	N/A
XBP1 H_2O_2Lo	XBP1_Q2	N/A	N/A	N/A	N/A
ZAP1 YPD	ZAP1_Q6	N/A	N/A	N/A	N/A

The first column indicates the ChIP-chip or PBM data set from which a given gene set was derived. All genes with ChIP p-value < 10^{-2} were thereby assigned to the respective sets. Columns 2 to 6 show the corresponding top ranking PFMs. Red indicates matching matrices, yellow indicates co-regulators. For instance, GAL4 target genes are oftentimes co-regulated via MIG1 (see Figure 4.18, page 86 for an example) while LAC9 and GAL4 both interact with GAL80 and bind to similar motifs (Zenke et al., 1993). Overall results match well to what is detected by the z-score test.

associations and indicate that important biological information about regulating TFs can straightforwardly be obtained from the ranking of the PFMs for a given data set or conversely, ranking data sets for a given PFM. The hypergeometric test has, however, the clear advantage of not requiring the *a priori* definition of any cutoffs. The hypergeometric test statistic will therefore be used for subsequent analyses.

Table 6.4 – Top ranking ChIP-chip set according to the hypergeometric test

MATRIX	RANK				
	1	2	3	4	5
ABF_C	*in vitro*	YPD	SM	N/A	N/A
ABF_01	*in vitro*	YPD	SM	N/A	N/A
ADR1_01	32	N/A	N/A	N/A	N/A
CBF1_B	SM	YPD	N/A	N/A	N/A
GAL4_01	raffinose	galactose	YPD	MIG1 *in vitro*	N/A
GAL4_C	raffinose	galactose	YPD	N/A	N/A
GCN4_01	SM	rapamycin	YPD	N/A	LEU3 SM
GCN4_C	rapamycin	SM	YPD	LEU3 SM	N/A
GCR1_01	YPD	N/A	N/A	N/A	N/A
GCR_B	YPD	N/A	N/A	N/A	N/A
HAC1_Q2	39	N/A	N/A	N/A	N/A
HAP1_B	YPD	N/A	N/A	HAP4 YPD	N/A
HAP234_01	HAP4 YPD	rapamycin	HAP2 YPD	HAP3 YPD	HAP1 YPD
HSF_04	H_2O_2Lo	H_2O_2Hi	MSN4 H_2O_2Hi	MSN2 Acid	MSN2 H_2O_2Lo
LEU3_B	SM	N/A	N/A	YPD	N/A
MATA1_01	105	N/A	N/A	N/A	N/A
MCM1_01	α Factor	YPD	STE12 but 14	STE12 but 90	N/A
MCM1_02	α Factor	YPD	STE12 but 14	N/A	N/A
MIG1_01	*in vitro*	GAL4 raffinose	YPD	N/A	N/A
STRE_B	MSN2 H_2O_2Hi	MSN4 H_2O_2Hi	MSN2 H_2O_2Lo	N/A	MSN2 acid
STRE_01	MSN4 H_2O_2Hi	MSN2 H_2O_2Hi	N/A	MSN2 H_2O_2Lo	N/A
PDR3_Q2	41	N/A	N/A	N/A	N/A
PHO4_01	N/A	phosphate	N/A	N/A	N/A
ROX1_Q6	30	N/A	N/A	N/A	N/A
RAP1_C	*in vitro*	YPD	SM	FHL1 SM	FHL rapamycin
RCS1_Q2	H2O2Hi	SM	N/A	N/A	N/A
REB1_B	N/A	YPD	H_2O_2Hi	N/A	N/A
STE12_Q4	α Factor	butanol 14	butanol 90	YPD	MCM1 YPD
XBP1_Q2	H_2O_2Lo	N/A	N/A	N/A	N/A
ZAP1_Q6	YPD	N/A	N/A	N/A	N/A

Columns 2 to 6 denote which ChIP data set was most significantly enriched with predicted targets of the PFM indicated in column 1. Red highlights matching TF-data set pairs; yellow indicates sets corresponding to known co-regulating TFs. Importantly, the ranking of the conditions is in agreement with what is known about the primary activity of the TFs. For instance, GAL4 is primarily active in galactose and raffinose containing medium where it is five fold higher expressed than in rich medium (Frolova et al., 1999). Similarly, LEU3 an activator of amino acid biosynthesis is active mainly in the absence of amino acids in the cell culture medium (SM condition, Wang et al., 1999). Non-matching associations are denoted as N/A.

Comparison to results from ROC curve AUCs

In the following the results obtained by PASTAA are compared to what is achieved by a ROC curve analysis. As shown in Chapter 4 and 5, ROC curves are a simple and fast to compute measure for the enrichment of true TF targets among the genes top ranking according to affinity. The ranking of the chip data sets for the different PFMs according to ROC curve AUCs is shown in Table 6.5 again based on defining all intergenic region with ChIP-chip p-value < 10^{-2} as. While about 50% of the top ranking associations match the expectation a considerable number of TF-data set pairs are not recovered by this procedure as compared to the z-score and hypergeometric tests. In addition, only a few co-factors are identified. Among the data sets not correctly recovered by ranking according to ROC curve AUCs are primarily those, for which the TRAP analysis in Chapter 4 did not yield significant correlation between affinity predictions and *R/G* ratios. These data sets likely constitute cases where only a weak enrichment is present. In addition, also for GAL4_01 and GAL4_C no meaningful associations are recovered by ROC curve AUCs. This is likely due to only a small number of genes having ChIP-chip binding p-value < 10^{-2} for the GAL4 data set. In turn, random rankings tend to yield spuriously large ROC curve AUCs in cases with only a small number of true positives. This also highlights the fact that any analysis based on ROC curves naturally requires the definition of a suitable cutoff separating genes into true positives and true negatives.

6.3.2 Validation on individual vertebrate genes

To test PASTAA's applicability to higher eukaryotes first its ability to detect the association between single genes and their regulating TFs has been investigated. This represents a continuation of the analysis shown in Table 5.2 where TRAP affinities were used to rank PFMs for a number of known SRF target genes. The present study thus allows to directly compare the simple TRAP ranking and the ranking provided by PASTAA using the hypergeometric test. The analysis is extended to a number of other known autoregulatory transcription factors each of which binds directly to its own promoter and thereby activates or represses its transcription. For this application a given input gene set supplied to the hypergeometric test consists only of the autoregulatory gene itself. The cutoff on the putative target genes is thus automatically chosen in such a way that the set contains only the genes with affinity ≥ to the affinity of the input gene itself. For the case of a single input gene the same PFM ranking could be obtained also by considerably simpler measures. Nevertheless, when using 1kb long promoters for computing the affinities for all 593 vertebrate TRANSFAC matrices, PASTAA finds for the four known SRF targets used in Section 5.3 (SRF, EGR1, ACT1, EGR3) only SRF matrices among the top four PFMs (see Table 6.6). In addition, for

Table 6.5 – Top ranking ChIP-chip data sets according to ROC curve AUCs

MATRIX	RANK				
	1	2	3	4	5
ABF1_01	YPD	*in vitro*	N/A	N/A	N/A
ABF_C	YPD	*in vitro*	N/A	N/A	N/A
ADR1_01	N/A	N/A	N/A	N/A	N/A
CBF1_B	YPD	SM	N/A	N/A	N/A
GAL4_01	N/A	N/A	N/A	N/A	N/A
GAL4_C	N/A	N/A	N/A	N/A	N/A
GCN4_01	rapamycin	SM	YPD	N/A	N/A
GCN4_C	rapamycin	YPD	SM	N/A	LEU3 YPD
GCR1_01	N/A	N/A	N/A	N/A	N/A
GCR1_B	N/A	N/A	YPD	N/A	N/A
HAC1_Q2	N/A	N/A	N/A	N/A	N/A
HAP1_B	YPD	N/A	N/A	N/A	N/A
HAP234_01	HAP4 YPD	HAP3 YPD	N/A	N/A	N/A
HSF_04	H2O2Lo	H2O2Hi	N/A	N/A	MSN2 acid
LEU3_B	N/A	N/A	N/A	SM	N/A
MATA1_01	N/A	N/A	N/A	N/A	N/A
MCM1_01	α Factor	YPD	STE12 but. 90	STE12 but. 14	N/A
MCM1_02	α Factor	N/A	YPD	N/A	N/A
MIG1_01	*in vitro*	N/A	N/A	N/A	N/A
STRE_01	MSN2 acid	MSN4 H2O2Hi	N/A	MSN2 H2O2Hi	N/A
STRE_B	MSN2 acid	N/A	MSN4 H2O2Hi	MSN2 H2O2Hi	HSF H2O2Lo
PDR3_Q2	N/A	N/A	N/A	N/A	N/A
PHO4_01	N/A	N/A	phosphate	N/A	N/A
ROX1_Q6	N/A	N/A	N/A	N/A	N/A
RAP1_C	YPD	*in vitro*	SM	FHL1 YPD	FHL1 rapa.
RCS1_Q2	N/A	H2O2Hi	N/A	N/A	N/A
REB1_B	N/A	YPD	H2O2Hi	N/A	N/A
STE12_Q4	butanol 14	α Factor	YPD	butanol 90	MCM1 YPD
XBP1_Q2	N/A	H2O2Lo	N/A	N/A	N/A
ZAP1_Q6	N/A	N/A	N/A	N/A	N/A

Columns 2 to 6 denote which ChIP data set yielded largest ROC curve AUCs for the PFM indicated in column 1. Red highlights matching TF-data set pairs; yellow indicates sets that correspond to known co-regulating TFs. In comparison to the hypergeometric and z-score test less ChIP-chip data sets are correctly identified. This applies in particular to sets with few ChIP-chip target genes such as in case of GAL4.

the autoregulatory gene CRX (Nishida et al., 2003) the corresponding matrix CRX_Q4 is ranked at position six. Finally, for the autoregulatory gene E2F2 (Neuman et al., 1994) all top ten PFMs correspond to matrices representing alternative E2F motifs. Together these results demonstrate that PASTAA improves the simple ranking according to affinities, which yielded among the top ranking TFs for the above genes a considerable fraction of likely spurious factors (compare results to Table 5.2).

Table 6.7 shows the results of this analysis for an extended group of autoregulatory genes in dependence on the choice of promoters used to compute the affinities. When using 200 bp proximal promoters half of these autoregulatory loops are detected with the corresponding PFMs ranking on top. In addition, for the CRX gene its PFM is now ranked at position 3. When extending the promoters to 500bps all but one autoregulatory loop are successfully detected (see Table 6.7). Additionally, among the top ten PFMs ranked for each of the autoregulatory genes PASTAA also finds other known regulators of these genes. This is true also for Irf1, the only case where the autoregulatory loop was not recovered. Instead NFKB and several STATs are found, which fits well to experimental findings (Luo et al., 2000; Wei et al., 2006). Another example are the genes HNF1 and HNF4 where PASTAA predicts as the top regulators HNF4 and HNF1, respectively, indicating cross regulation between the two factors. Together these findings indicate that autoregulatory binding signals tend to reside within proximal promoters and demonstrate that PASTAA is well suited for detecting vertebrate TFs associated with single genes.

Table 6.6 – Top ranking PFMs for several genes with known regulators

SRF	EGR1	ACT1	EGR3	CRX	E2F2
SRF_Q4	**SRF_Q4**	**SRF_01**	**SRF_01**	**PAX6_Q2**	**E2F1_Q4_01**
SRF_Q5_01	**SRF_01**	**SRF_C**	**SRF_C**	UF1H3BETA_Q6	**E2F_Q3_01**
SRF_Q5_02	**SRF_C**	**SRF_Q4**	**SRF_Q4**	SMAD_Q6_01	**E2F_Q4**
SRF_C	**SRF_Q5_01**	**SRF_Q5_01**	**SRF_Q6**	OLF1_01	**E2F_Q6**
SRF_Q6	**SRF_Q5_02**	**SRF_Q5_02**	E4F1_Q6	LEF1_Q2_01	**E2F1DP1RB_01**
PBX_Q3	**SRF_Q6**	**SRF_Q6**	RP58_01	**CRX_Q4**	**E2F4DP1_01**
OCT1_Q6	AHRARNT_02	MAZR_01	**SRF_Q5_01**	MAZR_01	**E2F_02**
OCT_C	AHR_01	HEN1_01	**SRF_Q5_02**	MZF1_02	**E2F1_Q6**
OCT_Q6	ATF6_01	HEN1_02	E2F1DP1RB_01	PU1_Q6	**E2F1_Q6_01**
TFIIA_Q6	YY1_Q6_02	AP2_Q6	HIF1_Q3	EVI1_05	**E2F_Q3**

Top ranking PFMs for the six genes shown in Table 5.2. The genes SRF, EGR1 ACT1 and EGR3 are experimentally known targets for SRF. CRX possesses a known autoregulatory loop as does E2F. Matrix identifiers in red indicate matching PFMs.

Table 6.7 – Autoregulatory binding signals reside within proximal promoters

TF	Promoter size					
	-200	-300	-500	-500 to +500	1000	2000
SRF	1	1	1	1	1	1
MEF2	1	1	1	1	3	4
CRX	3	5	5	7	6	12
E2F	1	1	1	1	1	1
IRF1	N/A	N/A	N/A	N/A	N/A	N/A
NFKB	1	1	1	1	1	1
HES1	1	1	1	1	1	3
PIT1	N/A	N/A	4	4	11	N/A
HNF4	N/A	4	3	4	4	5
HNF1	N/A	15	20	N/A	24	N/A

The majority of autoregulatory binding signals are located within 200 bps upstream of the respective TSS. All but one loop are detected within 500 bps upstream of the TSS. The binding signals decay if the promoter is extended further. Numbers in red indicate a rank below 10 (out of 593 TRANSFAC matrices). N/A indicates that the autoregulatory gene was not among the top 1000 genes when ranked according to affinity for the corresponding PFM. The hypergeometric test was thus not able to assign a meaningful p-value to such cases.

6.3.3 PASTAA can account for ChIP data from vertebrates

To assess PASTAA's ability to detect an enrichment of verified TF targets in a large set of vertebrate genes I turned to the ChIP-chip data set for the three hepatic transcription factors HNF1, HNF4 and HNF6 (Odom et al., 2004). In this study the binding of each of the three factors to ~13.000 human promoters was measured. As for the yeast ChIP-chip study, the sequences spotted on the utilized microarrays were used to compute the binding affinities for each of the 593 vertebrate PFMs contained in TRANSFAC. To speed up the computation of the hypergeometric tests, only promoters bound by HNF1, HNF4 or HNF6 (according to the usual p-value cutoff of $< 10^{-3}$) were assigned to the corresponding HNF1, HNF4 and HNF6 input gene sets. As shown in Table 6.8, for the HNF1 and HNF4 gene sets PASTAA correctly finds the highest association for the PFMs corresponding to HNF1 and HNF4, respectively. For the HNF6 data set the single HNF6 matrix present in TRANSFAC is ranked at position five while four matrices with similar motives to that of HNF6 yield even higher association p-values. Notably, among the top ten PFMs with highest enrichment for the HNF1 gene set PASTAA also lists HNF4, HLF (hepatic leukaemia factor) and the liver specific factor C/EBPalpha (Chen et al., 2000) while for the HNF6 data set HNF1 and also C/EBP are

Table 6.8 – Top associated PFMs for the HNF1, HNF4 and HNF6 target gene sets

HNF1 ChIP-chip	**HNF4** ChIP-chip	**HNF6** ChIP-chip	**MYC** ChIP-PET
HNF1_Q6	**HNF4_Q6_01**	CDPCR1_01	**E2F1_Q3**
HNF1_01	**HNF4_01**	CLOX_01	**E2F_Q2**
HNF1_Q6_01	**HNF4_01_B**	CDP_02	**MYC_Q2**
HNF1_C	**HNF4_Q6**	CDPCR3HD_01	ETF_Q6
AR_02	STAF_02	**HNF6_Q6**	**E2F1_Q4**
HNF4_Q6_01	**HNF4_DR1_Q3**	PBX1_02	ZF5_01
AR_03	**COUPTF_Q6**	**HNF1_C**	**MYCMAX_B**
HLF_01	T3R_01	CDPCR3_01	CHCH_01
CEBP_Q2	**COUP_01**	CDP_01	AP2ALPHA_01
RORA1_01	STAT_01	E2F_Q3_01	ZF5_B

Columns 1-3 show the top ranking PFMs for the HNF1, HNF4 (COUP) and the HNF6 ChIP-chip data sets. Column 4 indicates the top 10 matrices for the cMYC ChIP-PET data set. Matching matrices are indicated in red while PFMs corresponding to co-regulating factors are indicated in yellow.

detected. These results indicate that the above liver specific factors share many of their target genes and thereby confirm the experimental findings made by Odom et al., (2004) as well as the results from the previous subsection regarding cross-regulation between HNF1 and HNF4.

Similarly to the HNF data sets I also analyzed the cMYC ChIP-PET data set from Zeller et al., 2006. As suggested by the experimenters, the 1093 PET clusters with more than two overlapping PETs were selected as input sets. The affinities were computed for the entire sequences spanning the clusters (average length 2121 bps). 10.000 sequences of length 2121 bps with random genomic start positions were selected as background set. As shown in the rightmost column of Table 6.8, PASTAA finds two cMYC matrices among the top ten PFMs and another MYC matrix at position 13. Interestingly, among the top matrices several instances of E2F, an important coregulator of Myc target genes (Ogawa et al., 2002), are detected.

Having demonstrated that PASTAA is well suited for detecting regulating TFs in both yeast and vertebrates I now turn to a detailed analysis of human and mouse promoters of tissue specific genes. PASTAA is thereby employed to detect the location of binding signals within the promoters and subsequently to predict TFs that play an important role in the gene regulation of a given tissue.

6.4 Analysis of tissue specific promoters

The identification of transcription factors involved in tissue specific gene regulation remains a challenging task for both experimentalists and bioinformaticians. Previous computational methods have been able to recover some well-known TF-tissue pairs, however, the success of these methods has been limited to a relatively small number of individual tissues and factors whereby different methods tend to recover complementary sets of associations. For instance, while a dedicated analysis of retinal genes uncovered an important function for CRX and NRL in this tissue (Quian et al., 2005) a detailed analysis of muscle and liver specific genes found a strong association for SRF and MEF2 with muscle and revealed an important role of HNF1, 4 and 6 in liver (Johansson et al., 2003). Here PASTAA is applied for the detection of a more comprehensive list of TFs involved in the regulation of tissue specific genes. In order to find a maximal number of functional TF-tissue associations the first aim was to select the promoter regions around the TSSs optimally enriched in tissue specific binding signals.

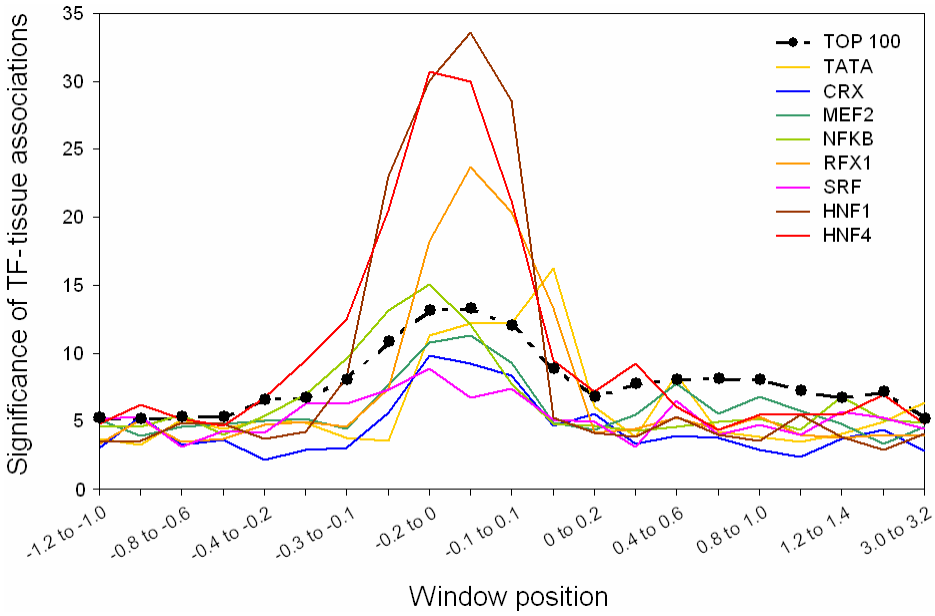
6.4.1 Detection of regulatory regions for tissue specific genes

When applying PASTAA to detect an enrichment of TF targets among a given input gene set the significance of the achieved hypergeometric p-values naturally depends on the definition of the putative promoter region used for computing $\langle N \rangle$. In the following this dependency was used to determine the promoter regions optimally enriched with tissue specific TF binding signals. This approach diverges from previous studies, which chose the promoter sequences in a rather *ad hoc* fashion to represent for instance 1kb upstream of each TSS (e.g. Zheng et al., 2003; Yu et al., 2006; Smith et al., 2007).

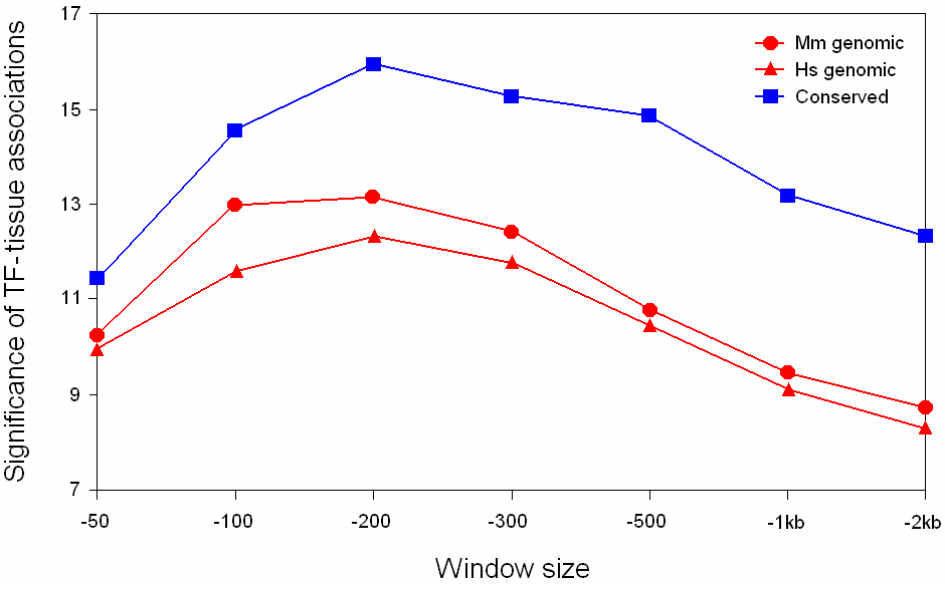
In order to find the optimal region I started by simultaneously shifting 200 bp windows in successive steps of 50 bps across all 26000 mouse promoters from Ensembl (version 31). All windows are thereby synchronized to start at the same distance x in respect to the TSS of their corresponding gene. For a given start position $\langle N \rangle$ is computed for all windows and all 593 TRANSFAC vertebrate PFMs. PASTAA is then used to evaluate the significance of the overlap between all combinations of 72 EST derived tissue categories and the target genes predicted based on the affinities from the windows starting at x. As objective test criterion for the suitability of a given promoter region the average -log p-values, μ, of the hundred most significant PFM-tissue associations obtained from a given window start position were evaluated. It is important to note that this averaging was not restricted to experimentally known TF-tissue associations but was performed without applying any prior knowledge. The

Figure 6.6 - 200bp proximal promoters yield most significant TF-tissue associations

a)



b)



a) Shown are the average -log hypergeometric p-values of the 100 most significant TF-tissue associations (black dotted line) as function of the location of the scanned 200 bp windows. The significance of the TF-tissue associations peaks when scanning a sequence from 0 to 200 bps upstream of each TSS. No strong signals are found for neighbouring windows. The trend is confirmed by the behaviour of the topmost association for many individual factors including SRF, MEF2 and NFKB. b) Enlarging or reducing the 200 bp core promoter region causes slow decay in the TF-tissue association signals for both human and mouse (red triangles and circles, respectively). Restricting the sequence space to evolutionary conserved mouse sequences increases the significance of the top 100 associations but does not change the location of the optimal promoter regions (blue squares).

implicit assumption made here is that the more significant the obtained p-values are the more enriched will the corresponding promoter region be with functional tissue specific binding signals. In contrast, windows containing primarily random sequence noise should not yield any significant associations between PFMs and tissues.

As shown in Figure 6.6a largest average -log p-values (μ = 13.2) are found for the two consecutive windows ranging from +50 to -150 and 0 to -200 bps in respect to the TSS. Average p-values from the neighboring windows quickly drop by several orders of magnitude and eventually level off at around μ = 5 for all windows outside of the region from +200 to -400. This trend reflects the behavior of many individual TFs such as the HNFs, SRF and MEF2. Notable exceptions are some GC rich motifs such as SP1, which achieve significant associations primarily with the brain tissue category also for windows located downstream of the TSSs (not shown).

Next, the effect of reducing or enlarging the 0 to -200 bp windows has been investigated. As shown in Figure 6.6b extending the promoter region further upstream causes a slow decay in the significance of the average p-values. This suggests that an increasing number of non-functional high affinity sites are being included. Reducing the size to 100 or 50 bps upstream of the TSS has a similarly detrimental effect indicating that in this case many functional sites are being removed. Finally, extending the region downstream of the TSS also causes a decline in the significance of the optimal p-values (μ = 11.8 for the region of +200 to -200 bps around the TSS) confirming the apparent lack of tissue specific binding signals downstream of the TSSs. A curve of nearly identical shape but somewhat smaller magnitude is obtained when using human promoters for the analysis. In contrast, restricting the sequence space to only conserved blocks between human and mouse greatly increases the significance of the tissue-TF associations without changing the location or size of the region producing the most significant results. The increased significance indicates that the high level of sequence conservation near the TSS (~39% at the TSS compared to ~9% for > 2 kb upstream) likely reflects the preservation of regulatory elements between mouse and human. From the peak in significance obtained from the 0 to -200 bp windows I conclude that this region yields the optimal trade-off between including functional sites and false positives. In the following this region of full genomic sequence will be referred to as the 200 bp proximal promoter (200PP).

6.4.2 TFs preferentially associated with EST derived tissue sets

Having found that 200 bp proximal promoters yield most significant TF-tissue associations I now investigate whether these regions indeed allow for the identification of experimentally known associations between tissue specific TFs and their corresponding tissues. To this end, all PFMs were ranked according to the significance of their association with a given tissue category obtained when computing the affinities for the 200PPs. Table 6.9 shows for a number of mouse EST tissue categories the most significantly associated matrices. In addition to some frequently identified tissue-TF pairs such as HNF1 and HNF4 for liver and kidney (Pennacchio et al., 2007, Smith et al., 2006, Yu et al., 2006) or SRF for heart and muscle (Wasserman et al., 1998) several other experimentally known associations - rarely or not revealed in previous studies - are detected. For instance, the two pancreatic transcription factors IPF1 (Insulin Promoter Factor 1, Ohlson et al., 1993) and PTF1 (Pancreas-specific Transcription Factor1, Roux et al., 1989) are listed among the top ten factors for pancreas together with HNF1 and two PFMs for HNF3 (HNF3_Q6, XFD3_01), which also play a known role in the regulation of pancreatic genes (Kaestner et al., 1999). The lung and thyroid specific factor TTF1 (Thyroid Transcription Factor 1, Kimura et al., 1999) is detected as the top ranking factor in the lung category and among the top ten factors in the thyroid category while the pituitary gland specific factor PIT1 (Pituitary-specific positive Transcription factor 1, Li et al., 1990) is detected near the top of the pituitary gland category. In retina the PFMs for the eye and pineal gland specific factor CRX (Cone Rod homeobox protein, Furukawa et al., 1999) and the eye specific factor CHX10 (Liu et al., 1994) rank at positions 2 and 3, respectively. In addition, PASTAA finds MAF to be associated with retina, which is known to form a dimer with the eye specific factor NRL (Sharon-Friling et al., 1998). Since TRANSFAC contains no PFM for NRL I have taken the matrix from Qian et al., (1994) and find a strong association with retina (rank 17). For the immune system related tissues spleen and thymus a comprehensive number of immune related factors including ETS1, IRF7 and NFKB are predicted. Finally, SRF, MEF2 and MTATA (muscle specific TATA) were observe as the top ranking PFMs for both the heart and muscle category. Similar results are obtained when using the small set of 13 hand picked muscle specific genes defined by Wassermann and Fickett (1998). To test whether the results derived from the EST categories depend on the incorporation of the gene set from Wasserman and Fickett these 13 genes were removed from the EST based lists. Surprisingly, the ranking of the factors for muscle and heart was not affected by this change indicating that a considerable number of genes assigned to the two tissue categories possess high affinity sites for SRF and MEF2 and that the detection of these associations is robust against limited changes in the input gene sets (data not shown).

Table 6.9 – Top ranking associations obtained for EST derived tissues

Intestine	Leukocyte	Pancreas	Diencephalon
HNF4_Q6_01	NFκB_01	*TATA_01*	NRSF_Q4
HNF4_01	NFκB65_01	PTF1β_Q6	EGR1_01
HNF4_DR1_Q3	ETS_Q6	RBPJK_Q4	ATF1_Q6
HNF4_01_B	NFKB_Q6	PR_Q2	*TFIII_Q6*
DR1_Q3	NFKB_Q6_01	HNF3_Q6_01	ATF_01
PPAR_DR1_Q2	NFKB_C	XFD3_01	NGFIC_01
COUP_DR1_Q6	CREL_01	E12_Q6	CACB._Q6
TATA_01	ELF1_Q6	IPF1_Q4_01	NRSE_B
HNF4α_Q6	ETS_Q4	HNF1_Q6	T3R_Q6
SRF_Q5_02	ETS1_B	E47_02	AP1_01
Kidney	**Spleen**	**Retina**	**Thyroid gland**
HNF1_01	IRF_Q6	GATA1_03	*TATA_01*
HNF1_Q6	ISRE_01	CRX_Q4	*TATA_C*
HNF1_Q6_01	ICSBP_Q6	CHX10_01	CHOP_01
HNF4_01	IRF_Q6_01	SREBP_Q3	SREBP_Q3
HNF1_C	NFκB_01	CHCH_01	MEF3_B
HNF4_01_B	NFKB_Q6_01	LRF_Q2	TTF1_Q6
DR1_Q3	IRF7_01	**VMAF_01**	SREBP1_02
HNF4_Q6_01	NFKB_Q6	RREB1_01	CEBPδ_Q6
HNF4_DR1_Q3	IRF1_01	AP4_Q5	NKX25_01
COUP_01	STAT3_01	CACB._Q6	TITF1_Q3
Striated muscle	**Thymus**	**Testis**	**Pituitary gland**
MEF2_02	ETS_Q6	RFX1_01	*TATA_01*
SRF_C	ETS_Q4	**VMYB_02**	**NGFIC_01**
SRF_01	ETS1_B	CREBP1CJUN_01	POU1F1_Q6
SRF_Q5_02	AML_Q6	RFX1_02	T3R_01
SRF_Q5_01	E2F_Q2	MIF1_01	**EGR2_01**
SRF_Q6	PPARG_02	EFC_Q6	MEIS1BHOXA9_02
MTATA_B	CBF_02	**CMYB_01**	*TATA_C*
RSRFC4_01	MYOD_Q6_01	VJUN_01	**ER_Q6**
MEF2_Q6_01	SEF1_C	TCF11_01	CHOP_01
MEF2_03	E2F1_Q4	RFX_Q6	DEAF1_02

The table shows for twelve tissues the ten most significant tissue-matrix associations obtained when computing ⟨N⟩ for the 200 bp proximal promoters. Tissue-TF pairs with strong support in the literature are indicated in red, associations with likely function are indicated in bold and TATA box motifs are indicated in italic.

Table 6.10 – Enrichment of TATA box motifs in the first 50 bps upstream of the TSS

Tissue	PFM				Lowest rank
artery	**TATA_01**	**TATA_C**	**MTATA_B**	**ATATA_B**	**1**
breast	**MTATA_B**	**TATA_01**			**1**
cartilage	**TATA_01**	**TBP_Q6**			**1**
cerebrum	ZF5_B	NRF1_Q6	KROX_Q6	SP1_Q6_01	56
diaphragm	**TATA_01**	**MTATA_B**	**TFIIA_Q6**		**1**
eye	**TATA_01**	AP2_Q6			**1**
heart	**MTATA_B**	**TATA_01**			**1**
intestine	**TATA_01**	**MTATA_B**	**TBP_Q6**	**TATA_C**	**1**
kidney	HNF4_01	**MTATA_B**	**TATA_01**	HNF4_Q6_01	**2**
leukocyte	ETS_Q6	PEA3_Q6	PU1_Q6	ELF1_Q6	**9**
liver	**TATA_01**	SRF_Q5_02	**TBP_Q6**	**ATATA_B**	**1**
lung	**TATA_01**	**TATA_C**			**1**
muscle	**MTATA_B**	**TATA_01**	AMEF2_Q6	MMEF2_Q6	**1**
olfactory epi.	CEBPG._Q6	GATA3_03	GATA1_05	EVI1_03	**7**
pancreas	**TATA_01**	**TBP_Q6**	**MTATA_B**	**TATA_C**	**1**
peri. nerve	**MTATA_B**	**TATA_01**			**1**
pineal gland	APOLYA_B	**TATA_01**			**2**
pituitary gland	**TATA_01**	**TBP_Q6**	**MTATA_B**	**TATA_C**	**1**
placenta	**TATA_01**	**MTATA_B**	**ATATA_B**	**TBP_Q6**	**1**
retina	GATA1_03	OTX_Q1	CRX_Q4		78
salivary gland	**MTATA_B**	**TATA_01**	**TATA_C**	**MMEF2_Q6**	**1**
skin	**TATA_01**	**TBP_Q6**	GZF1_01	**MTATA_B**	**1**
spinal cord	**MTATA_B**	**TATA_01**			**1**
spleen	**TATA_01**	ICSBP_Q6			**1**
stomach	**TBP_Q6**	**TATA_01**	**MTATA_B**	**TATA_C**	**1**
testis	VMYB_02	RFX1_01	RFX1_02	MIF1_01	72
thymus	ETS_Q6	PU1_Q6			112
thyroid gland	**TATA_01**	**TATA_C**	**MTATA_B**		**1**
tongue	**TATA_01**	**TATA_C**	**MTATA_B**	SRF_Q5_02	**1**
venae	**TATA_01**	**TATA_C**	**ATATA_B**		**1**
vesicular gland	**TBP_Q6**	CDXA_01	**TATA_01**	EVI1_03	**1**
whole brain	ETF_Q6	ZF5_01	E2F_Q2	ZF5_B	42

A strong enrichment of TATA box elements can be observed in most tissue categories when analyzing core promoters. The top four factors exceeding a hypergeometric p-value of 10^{-6} are shown for each tissue. Top ranks (out of 593) for any TATA box motif in a given tissue category are indicated in the last column. Obvious exceptions are cerebrum, retina, testis, thymus and whole brain (indicated in bold). Experimentally verified tissue-TF association are indicated in italic.

Tissue specific genes are associated with TATA box motifs

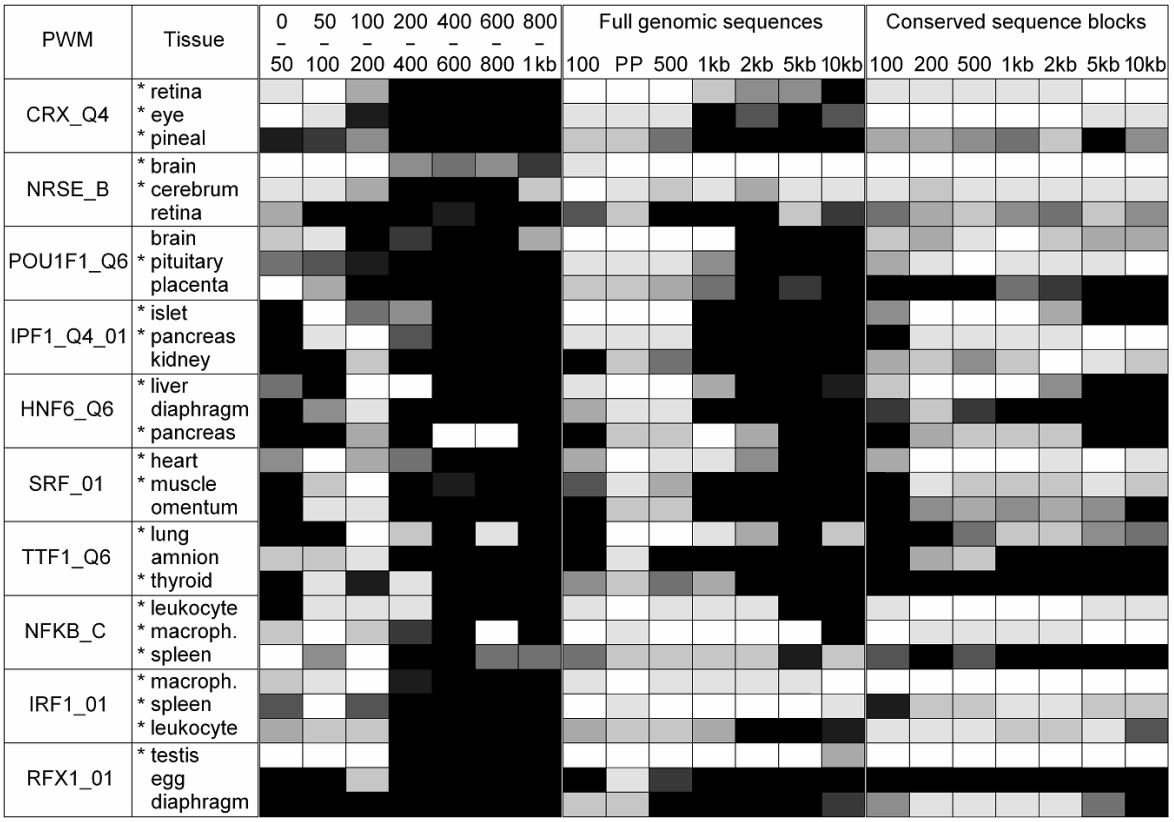
By using the 200PPs PASTAA detects TATA box motifs among the top ten associations in a variety of tissues (see Table 6.9). This trend becomes even more evident if smaller upstream regions are used to compute ⟨N⟩. In fact, TATA motifs are top ranking in the majority of tissue categories when limiting the promoters to only 50 bps upstream of the TSSs. As indicated in Figure 6.6a this enrichment goes in hand with a peak in the association p-values between the TATA box matrices and the tissue categories. Table 6.10 shows the top four PFMs exceeding a hypergeometric p-value of 10^{-6} for each tissue. The motifs TATA_01, TATA_C, ATATA_B (avian TATA box), MTATA_B (muscle specific TATA box), TFIIA_Q6 (general transcription factor IIA) and TBP_Q6 (TATA binding protein) are prominently ranked near the top of most categories. Exceptions are whole brain, cerebrum, retina, testis and thymus. In the case of thymus, testis and retina various specific TFs are ranked on top while in whole brain and cerebrum GC rich motifs such as ETF_Q6 and ZF5_01 are dominating. In contrast to TATA box motifs most experimentally known tissue-PFM associations become far less significant when considering only the first 50 bps upstream of the TSS. Together these findings indicate that regulatory motifs of tissue specific TFs tend to be located more than 50 bps upstream of the predicted TSSs and that aside from primarily neuronal tissues many tissue specific genes possess a TATA box. The latter observation is in agreement with the notion that the expression of highly regulated genes is often TATA-dependent whereas expression of housekeeping genes, which are expressed in all cell types, is TATA-independent (Yang et al., 2007).

6.4.3 Tissues preferentially associated with a given TF

As demonstrated in Section 6.3.1, the computed hypergeometric p-values can be used to reverse the search and identify gene groups most strongly associated with a given TF. In order to identify the top ranking tissues for a given TF and to assess how well this ranking agrees with experimental knowledge in the following I thus switched from a tissue centric to a TF centric view. The subsequent analysis was thereby performed using the association p-values as computed above for the 200PPs but this time ranking tissues instead of PFMs.

For a group of ten PFMs the three top ranking tissues are shown in the first two columns of Figure 6.7. The listed associations are in good agreement with the experimentally known tissue specific functions of the corresponding TFs. For several TFs two or more tissues in which the factor plays a known role compete for the top rank. For instance, HNF6 plays a role not only in liver but also in pancreas (Odom et al., 2004), CRX has been described as a retina and pineal gland specific factor (Furukawa et al., 2002) and TTF1 plays a dual role as thyroid and lung specific factor (Kimura et al., 1999).

Figure 6.7 – Dependence of tissue ranks on the promoter size

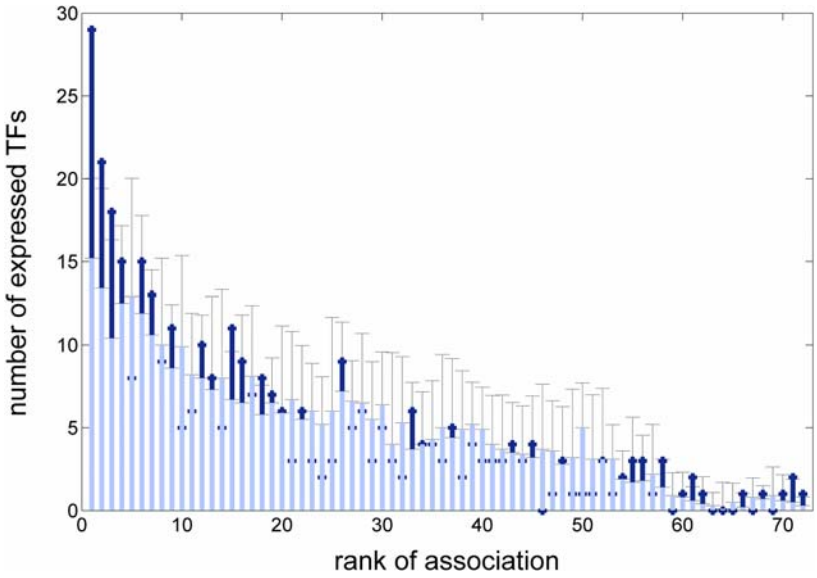


Ten PFMs and their top ranking tissues obtained when computing ⟨N⟩ for the 200PPs are shown in columns 1 and 2. Associations supported by literature are indicated by an asterisk. Grey shades symbolize the ranks of the tissues obtained when computing ⟨N⟩ for different promoter regions (white: rank 1; black: rank ≥ 10). Columns in the left panel correspond to regions of indicated range upstream of the TSS. Regions located between 0 and 400 bps thereby frequently confirm same TF-tissue associations. Columns in the central and rightmost panel correspond to ⟨N⟩ being computed for upstream regions of indicated size starting that the TSS, with considering the full genomic sequence (central panel) or only the conserved sequence blocks (right panel). For full genomic sequences the detection of experimentally confirmed associations is robust against extending the promoters to ≈ 500 bps upstream of the TSSs. In contrast, using phylogenetic footprinting yields oftentimes identical tissue rankings even when extending promoters to > 10kb.

TFs tend to be themselves over-expressed in their top ranking tissues

To assess how meaningful the top ranking tissue associations are I first analyzed the expression patterns of the TFs themselves. The underlying assumption is that a TF specifically expressed in a certain tissue is likely to play a regulatory function in the very same tissue. In turn a TF should be over-expressed more frequently among its top-ranking

Figure 6.8 – TFs are over expressed in their top ranking tissues



Tissues top ranking for a given TF express the factor more often than expected, while bottom ranking tissues express the TF equally or less often than expected. This is indicated by the height of bins representing the number of TFs expressed in the associated tissue of given rank based on the real sequence data (dark blue) or on the results obtained from random sequence sets (light blue). Error bars show the 95% confidence interval for the results obtained from the random sequence sets. The enrichment is particularly significant for the first three bins corresponding to all the three top ranking TF-tissue associations (p-value of enrichment for bins 1-3 combined: 2.2×10^{-12}). The general trend in the light blue bins indicates the technical bias caused by the dependency of the hypergeometric tests on the number of ESTs in the tissue categories.

tissues rather than among randomly assigned tissues. In the entire data set there are 352 TF-tissue associations where the TF is specifically expressed in the corresponding tissue (EST cluster p-value < 10^{-6} for the respective TF). In 29 of these cases the tissue is indeed top ranking for the TF. This constitutes a twofold increase (p-value for enrichment: 1.3×10^{-6}) over what would be expected based on the randomization procedure outlined in Section 6.2.3. In 21 cases the tissue is ranked second (1.6 fold increase, p-value 0.019) and in 17 cases third to top (1.7 fold increase, p-value 0.017). Over all 72 possible ranks (corresponding to all 72 tissues) a clear trend exists for the higher ranking tissues to express the corresponding TFs more often than expected while lower ranking tissues tend to express the TFs as often as or less frequently than expected (Figure 6.8). These results are stable against removing groups of tissues from the analysis (for instance all immune related tissues) or performing the analysis for all 593 PFMs instead of the reduced set of corresponding TFs.

While these findings support a considerable number of predicted TF-tissue associations this verification method fails for factors such as SRF and HNF1, which are broadly expressed despite their known tissue specific activities, or for factors such as PTF1, which do not have enough support by EST or GNF data to validate their expression patterns. To confirm such TF-tissue associations I performed a manual Pubmed search looking for strong evidence for the involvement of the TF in the regulation of the respective tissue. Mapping the findings back to the matrices both validation methods together strongly support 282 out of the 1779 (three times 593 PFMs) three top ranking PFM-tissue associations (133 through expression data and 149 through literature). In the following, these cases will be considered as a test set of verified associations.

Exemplary analysis of eye specific target genes for CHX10 and CRX

As a detailed example of two verified associations between TFs and a tissue category I investigated more closely the link between the homeobox transcription factors CRX and CHX10 and the retina specific gene set. As shown in Table 6.9, both TFs are found to be associated with the eye and retina category. In addition, both factors are themselves significantly expressed in terminally differentiated retina and are known to be involved in the development of the eye. CRX is a known transactivator of retina specific genes (Furukawa et al., 2002) whereas CHX10 has been proposed to act as a transcriptional repressor for a number of eye specific genes including several CRX targets (Dorval et al., 2006). While the binding motifs of the two factors share the homeobox core (ATTA) the flanking bases are clearly distinct from each other leading to the proposal that the two factors bind primarily distinct sites (Dorval et al., 2006). The eye specific genes considered as targets for the two TFs by PASTAA are shown in Table 6.11. Especially the gene encoding the transcription factor *Mab21l1*, which is also required for proper eye development, is ranked among the top targets for both factors. *Mab21l1* has been placed in the same developmental pathway as CHX10 (Yamada et al., 2004) with CHX10 being severely down regulated in *Mab21l1* knockout mice. Figure 6.9 illustrates the *Mab21l1* locus with the predicted DNA binding affinities according to TRAP and predicted binding sites according to the balanced cutoff method (Rahmann et al., 2003) for CRX and CHX10. In agreement with the observation of Dorval et al., (2006) the majority of predicted binding sites are distinct for the two factors. However, the sites with highest predicted affinity for CRX and CHX10 (located near the TSS of *Mab21l1*) are overlapping. This suggests competitive binding of CRX and CHX10 at the *Mab21l1* locus. Furthermore, with *Mab21l1* being a positive regulator of CHX10 expression and CHX10 acting as a transcriptional repressor it implies the presence of a negative feedback loop for *Mab21l1* in the absence of CRX. While this analysis indicates the amount of detailed biological information that can be obtained when combining expression data with

transcription factor affinity predictions we now return to the large scale analysis of tissue specific promoters.

Table 6.11 – Eye specific targets for CHX10 and CRX in mouse

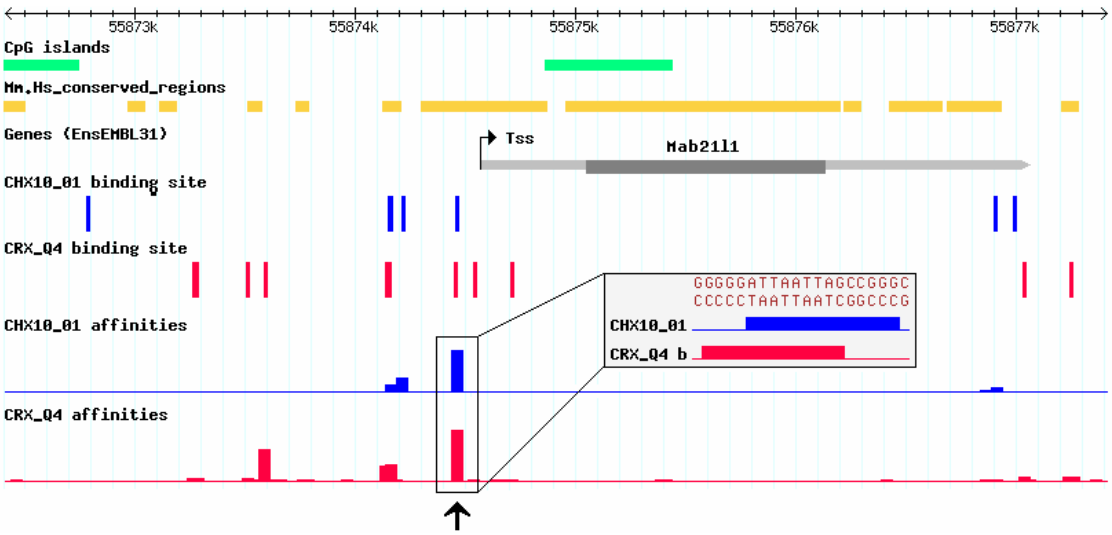
MATRIX	Retina target	EST rank	Affy rank
CHX10_01	Spacrcan	113	20
	Lpgds	58	43
	Mab21l1	**77**	**46**
	Arr3	**55**	**80**
	Abc10	**24**	**86**
CRX_Q4	Crx	18	5
	Mab21l1	**77**	**25**
	Nr1f2	100	76
	Vtn	98	106
	Gng1	85	109
	Brown	43	246
	Prph2	10	251
	Bc027072	78	297
	Lrrc21	94	354
	Gcap1	22	398
	Gnb5	108	408
	Car	11	422
	Cplx4	83	441
	Rs1h	26	488
	Arr3	**55**	**595**
	Samd7	158	685
	Abc10	**24**	**818**
	Cabp4	109	876

First two columns indicate the PFM and the corresponding eye specific target genes as predicted by PASTAA. Column 3 shows the expression rank in the EST data set. The last column shows the corresponding affinity rank of the target gene in respect to the entire set of 26609 mouse genes. Indicated in red are target genes shared between the two TFs. Most notably, the gene *Mab21l1* is ranked near the top for both TFs while the CRX gene is ranked among the top CRX targets.

6.4.4 Distribution of functional binding signals across promoters

In the following the distribution of binding signals for the set of 282 top ranking TF-tissue associations (page 136) among the corresponding promoters will be investigated. In this context it is important to remember that the selection of the 200 bp proximal promoters in Figure 6.6 was simply based on the average significance over the top 100 associations without restricting the search to experimentally verified TF-tissue associations. The analysis of the verified associations is subsequently extended to explore how robust the top ranks are against changing the promoter definition and how well the ranks are conserved between human and mouse.

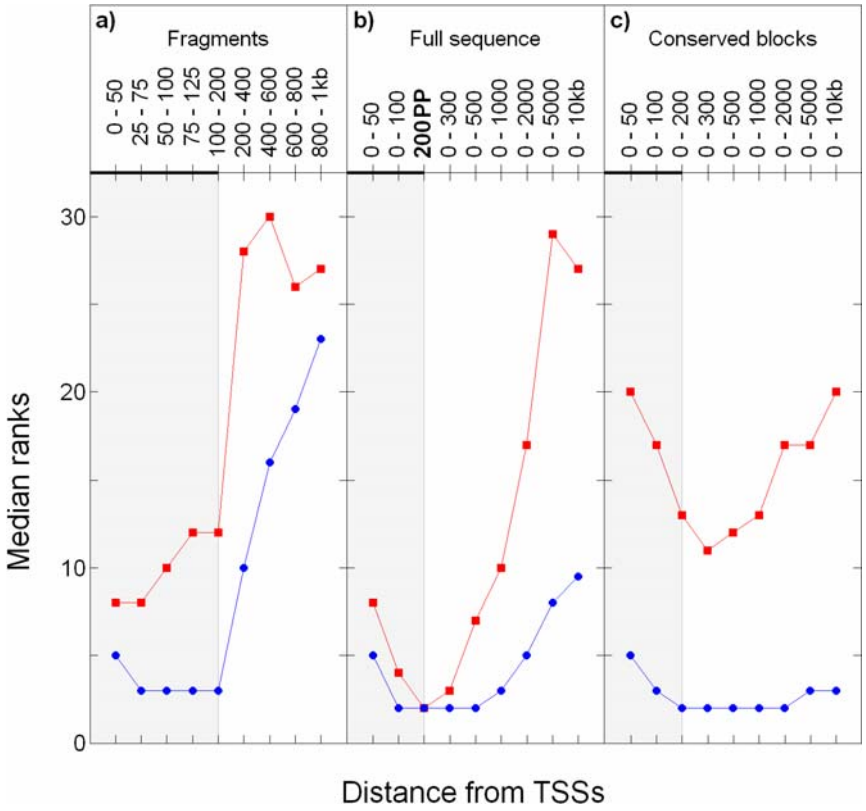
Figure 6.9 – Visualization of the genomic region around the retina gene *Mab21l1*



Predicted binding behavior of CRX and CHX10 across the *Mab21l1* gene locus. CpG islands, conserved bocks and the *Mab21l1* ORF are indicated on top. Binding sites for CHX10 and CRX, as predicted by the balanced cutoff method, are indicated below. TF affinities as predicted by TRAP are indicated at the bottom. As highlighted by a black arrow highest predicted affinities lie in a conserved block within the proximal promoter of *Mab21l1* while predicted binding sites are more broadly distributed across the gene locus. The inset shows that the high affinity sites near the TSS overlap, suggesting competitive binding between CRX and CHX10.

As a first test, the tendency of different promoter fragments to confer same tissue ranks as the 200PPs was analyze. To this end, the ranks of the 282 confirmed TF-tissue associations for alternative promoter regions were compute. As a global measure of how well a region agrees with the 200PPs next the median over the ranks for all the 282 associations (in Figure 6.7 this would correspond to taking the median over one of the grey shade columns) was taken. A region leading to the same results as the 200PPs thereby got a median rank of 2 while random sequences obtained a median rank of about 36 (as there are 72 tissues). The results of this *in silico* promoter bashing analysis are shown in Figure 6.10a. The 50 bp long fragments located between 25 and 200 bps upstream of the TSS all yielded a median rank of 3 demonstrating strong independent support for the ranks obtained from the entire 200PPs. The fragments ranging from 0 to 50 bps obtained a median rank of 5 reconfirming the notion that the verified TF-tissue associations become less significant when limiting the region to only the core promoter (see Table 6.10 for comparison). Remarkably, the fragments from 200 to 400 bps upstream of the TSS, which were not included in the 200PPs, yielded a median rank of 10 indicating that this region has a considerable tendency to contain tissue

Figure 6.10 – Stability of ranks in respect to changing the promoter definition



The median rank over the set of 282 verified TF-tissue associations remains near the minimum of 2 for fragments included in the 200PPs (blue circles, panel a). For the verified associations also adjacent upstream fragments tend to assign the same rank as the 200PPs. In contrast, much larger median ranks are obtained for a group of 45 PFMs which are not expected to have any real tissue specificity (red squares). Panel b) shows that the ranks obtained for the verified TF-tissue associations only slowly deteriorate when the proximal promoters are extended while the ranks for the associations in the control set are very sensitive to changing the sequence space. Including phylogenetic footprinting allows extending the core promoters much further before the median ranks of verified associations deteriorate (panel c). In contrast the sequence signals for the control set seem to be only marginally conserved between mouse and human as is indicated by their large median ranks even when using the region of the 200PPs.

specific sequence signals in agreement with those found in the 200 bp proximal promoters. Analyzing fragments further upstream of the TSSs resulted in increasingly higher median ranks emphasizing that tissue specific sequence signals tend to be located near the TSSs. As a control, this analysis was also performed for the set of 45 PFMs (respectively the 135 corresponding top three associations) representing those TFs which are least likely to display any real preference for any of the 72 tissue categories. This group of TFs includes general

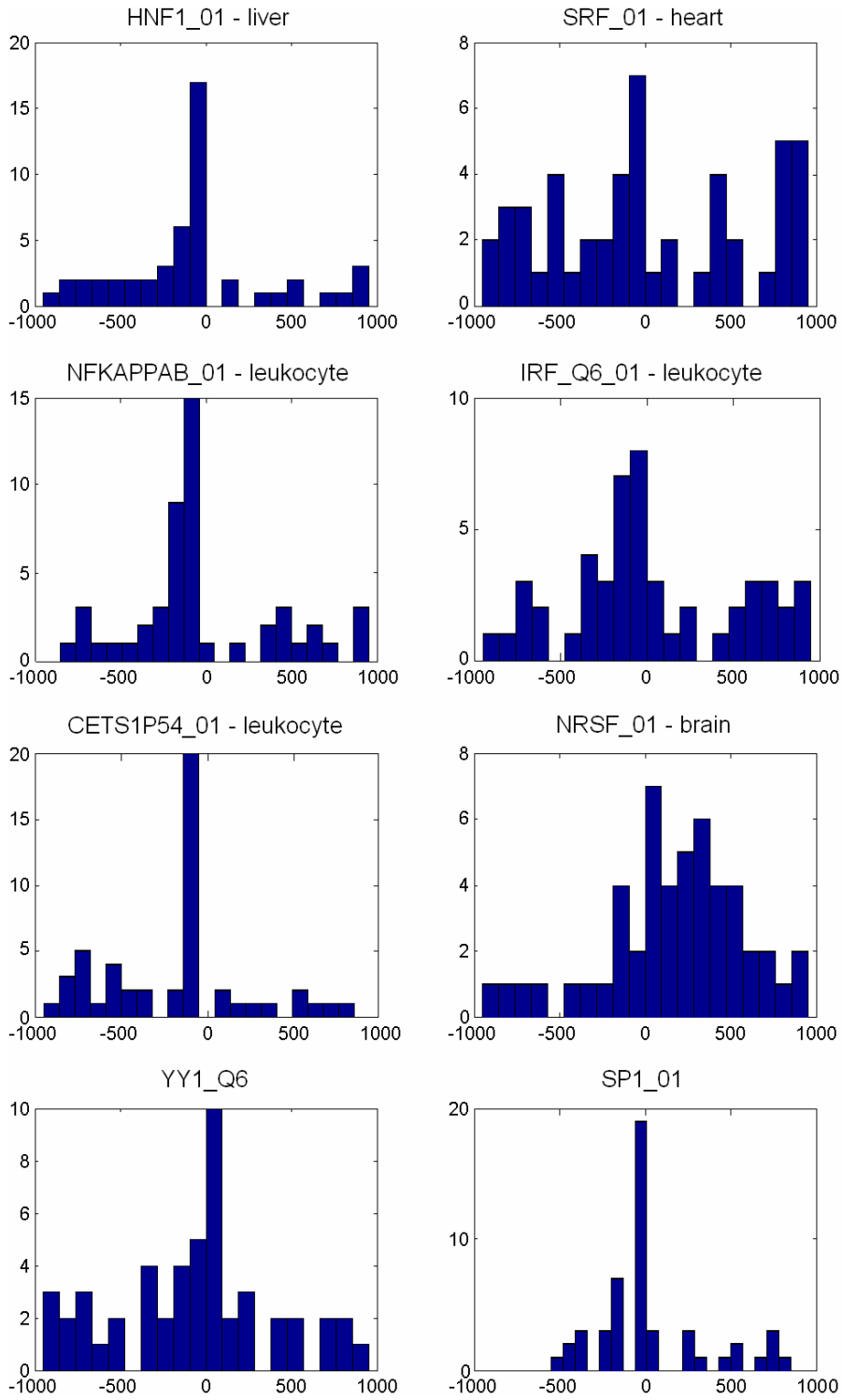
TFs such as TATA box binding protein, SP1, heat shock factor and P53, but also viral factors such as papilloma virus regulator E2 or the lentiviral TATA upstream element. Indeed, when using these PFMs there was a much weaker tendency, even for the fragments within the region from 0 to 200 bps, to support the tissue ranks assigned by the 200PPs while any fragments located further upstream produced seemingly random median ranks (red curves in Figure 6.10a).

Windows with maximal affinity are located near the TSS

All of the above evidence supporting the notion that functional binding signals are located primarily within 200PPs has been based on the significance of the found TF-tissue associations. TRAP affinity predictions, however, offer the possibility to confirm these findings by locating the region preferentially containing the high affinity sites for a given tissue specific TF. To address this question I used again the sliding window approach introduced in Section 5.3. For the current analysis a window of length 100bps was shifted in steps of 100bps across the promoter (± 1kb around the respective TSS) of a given tissue specific gene. The affinity of each window was computed and the location of the window with highest affinity for the TF was evaluated. The histogram in Figure 6.11 shows the preferential location of the windows with largest affinity for the 50 tissue specific genes with highest overall affinity for the corresponding TF. For instance, for the 50 genes with highest affinity for HNF1 and specific expression in liver (EST p-value < 10^{-6}) there exists a clear trend for the windows with largest affinity to be located near the TSS (about 50% are located within the 200PPs). A similar trend is observed for many of the TFs with verified tissue specific association. A notable exception is presented by the neuronal-gene repressor NRSF. The high affinity sites of this factor show a tendency to cluster downstream of the TSSs. In general, the preferential location of highest affinities close to the TSS is observed only in the appropriate tissue of the given tissue specific TF. For instance, no preference for the 200PPs is observed for HNF1 when analysing brain specific genes (not shown). In contrast, the windows with largest affinity for the two general TFs SP1 and YY1 cluster near the TSS across all tissues in accordance with experimental findings (Hui et al., 2007).

Together these findings represent strong additional support for the identified top ranking TF-tissue associations and further underline the tendency of tissue specific binding signals to accumulate within the 200PPs.

Figure 6.11 – Location of windows with maximal affinity around the TSSs



Histograms of the location of the max affinity windows within a region of ± 1kb around the TSSs for several PFM-tissue combinations. For the top six factors the analysis was carried out in their most significantly associated tissues. The two general factors YY1 and SP1 were tested across all tissues. Strong peaks in the histogram indicate evolutionary pressure to keep the sites near the TSSs.

TF-tissue ranks are robust against changes in promoter size and applying phylogenetic footprinting

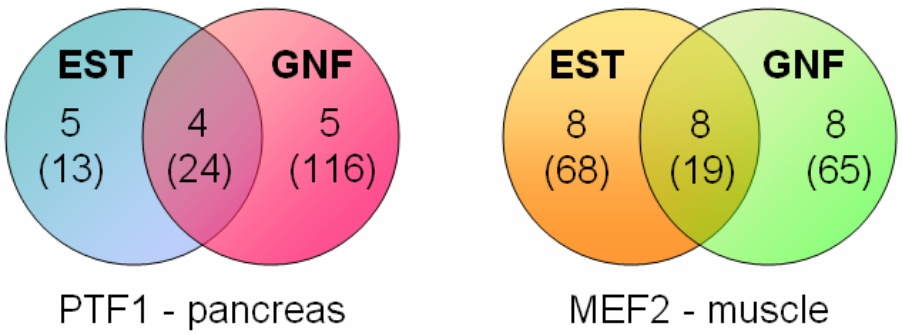
Having confirmed that high affinity sites for tissue specific TFs tend to accumulate within the 200PPs I now assessed the robustness of the computed ranks for the confirmed TF-tissue associations against enlarging the promoter regions. That is, how much sequence noise can be included before the ranks of the confirmed associations deteriorate? For the matrices shown in Figure 6.7 the tissue rankings are relatively stable against enlarging the genomic sequences to a size of ≈ 500 bps upstream of the TSSs (limited change of grey shades in the central panel of Figure 6.7). However, extending them to > 1kb causes many of the associations to obtain increasingly larger ranks. This signal decay can presumably be counteracted by applying phylogenetic footprinting and indeed, when limiting the sequence space to only conserved blocks most of the confirmed TF-tissue associations remain top ranking even if regions >10 kb are scanned (rightmost panel of Figure 6.7). One of the few exceptions is the associations of TTF1 with thyroid gland, which appears to be not conserved.

For a general assessment of the robustness of the PASTAA analysis against sequence noise I again determined the change in median ranks obtained for the verified set of 282 TF-tissue associations or the control set of general TFs (Figure 6.10b and c, respectively). In case of full genomic sequences median ranks remain at the best possible value of 2 up to a promoter size of 500 bp while further enlarging the promoters causes the median ranks to slowly increase. When applying phylogenetic footprinting the median ranks stay nearly unchanged even when extending the promoters to 10kb upstream of the TSSs. In addition, using conservation allows in some cases the detection of associations not found when using the full genomic sequences. For instance, while the known association between MYOD and muscle is ranked only at position 14 when using the full genomic 200PPs it is ranked on top when restricting the sequence space to the evolutionary conserved blocks. Another interesting change in ranks accompanying the application of phylogenetic footprinting is observed for the eye and pineal gland specific factor OTX2 (Nishida et al., 2003). Here the top ranking association switches from the pineal gland category to retina when applying conservation. Together these findings indicate that the majority of verified TF-tissue associations are highly conserved between mouse and human and that sequence noise in the upstream regions is effectively filtered out by applying phylogenetic footprinting. In contrast, when performing the analysis for the control set of 45 PFMs corresponding to TFs with tissue independent function median ranks appear highly sensitive to enlarging the sequence space and change greatly when applying phylogenetic footprinting, indicating that the top ranking TF-tissue associations assigned to these factors are, as expected, rather meaningless.

6.4.5 Overlap between GNF and EST data

For comparative purposes and to validate the above findings made based on EST derived groups of tissue specific genes I next switched to the comprehensive GNF microarray tissue expression data from Novartis (Su et al., 2004). As shown in Table 6.12, when using the tissue categories derived from the GNF data set, nearly identical top ranking associations are found for most of the corresponding tissue categories including retina, pancreas and spleen. Exceptions are thymus (not shown), thyroid gland and pituitary gland where the appropriate TFs could not be detected among the top ten matrices. Intriguingly, also the promoter regions yielding the most significant results are identical to those obtained from EST based tissues thereby strongly validating the previous findings. A possible reason for the similarity of the detected tissue-TF associations could be high correspondence between the expression patterns obtained from EST and GNF data. To test this possibility I analyzed the overlap among the top 100 genes in each of the corresponding tissue categories and found that in nearly all cases less than 50% of the genes overlap. This is in accordance with previous studies, which showed that microarray and EST data often predict different gene expression profiles (Munoz et al., 2004). To investigate in more detail how PASTAA is nevertheless able to detect identical associations for corresponding EST and microarray tissue categories I evaluated the overlap between EST and GNF data only among the target genes of a given TF as predicted by PASTAA. The results of this analysis are shown exemplarily in Figure 6.12 for the pancreatic factor PTF1 and the muscle specific factor

Figure 6.12 – Overlap between GNF and EST tissue assignments



The left Venn diagram shows the overlap of PTF1BETA_Q6 targets (as chosen by PASTAA) among the genes in the pancreas categories derived from EST and GNF data, respectively (top numbers) and the overlap between the genes considered to be pancreas specific according to each data set (numbers in parenthesis). The right panel illustrates the situation for MEF2_02 and the genes assigned to the skeletal muscle categories. For this factor, the intersection between EST and GNF data shows the most significant enrichment with predicted MEF2 targets (8 out of 19 genes).

Table 6.12 – Top ranking GNF tissue-PFM associations obtained from 200PPs

Small intestine	Lymphnode	Heart	Pancreas	Hypothalamus
HNF4_01	ETS_Q6	MEF2_02	CDXA_02	KROX_Q6
HNF4_Q6_01	ELF1_Q6	MEF2_Q6_01	PTF1BETA_Q6	HIC1_03
HNF4_DR1_Q3	IRF_Q6	MEF2_03	FOXO1_02	EGR_Q6
DR1_Q3	PU1_Q6	RSRFC4_01	NFAT_Q4_01	E2F_Q2
HNF4_01_B	ICSBP_Q6	RSRFC4_Q2	IPF1_Q4_01	ZF5_B
HNF4_Q6	NFkB_01	SRF_C	CEBP_01	ZF5_01
PPAR_DR1_Q2	ETS1_B	SRF_01	CDXA_01	ATF1_Q6
HNF4α_Q6	ISRE_01	KAISO_01	HNF3B_01	NRSF_Q4
COUP_DR1_Q6	NFKB_Q6_01	ATF_B	NRL_HAND	AHR_Q5
HNF1_01	NFKB_Q6	CREB_Q4	TATA_01	NRSF_01
Kidney	**Spleen**	**Skeletal muscle**	**Retina**	**Pituitary**
HNF1_01	**PTF1BETA_Q6**	MEF2_02	GATA1_03	ATF6_01
HNF1_Q6	ELF1_Q6	MEF2_01	CRX_Q4	EGR2_01
HNF1_Q6_01	ICSBP_Q6	AP4_01	AP1_C	STRA13_01
HNF4_01	IRF2_01	MEF2_03	**CMAF_01**	AHRARNT_01
HNF1_C	IRF1_01	MYOD_Q6_01	**VMAF_01**	ELK1_02
HNF4_01_B	STAT6_02	MYOGENIN_Q6	SPZ1_01	NFY_Q6_01
HNF4_Q6_01	IRF_Q6	HMEF2_Q6	CAAT_C	AP2GAMMA_01
HNF4_DR1_Q3	IRF_Q6_01	RSRFC4_Q2	KROX_Q6	GABP_B
DR1_Q3	ISRE_01	RSRFC4_01	TATA_01	CETS1P54_03
PPAR_DR1_Q2	TFE_Q6	KAISO_01	PITX2_Q2	STAF_01
Liver	**Cd4+Tcell**	**Testis**	**Thyroid**	**Amygdala**
HNF1_01	CETS1P54_02	VMYB_02	CDXA_02	ZF5_01
HNF4_Q6_01	CETS1P54_01	RFX1_02	VMAF_01	ETF_Q6
HNF1_Q6	ETS_Q6	RFX1_01	CDXA_01	E2F_Q2
HNF1_Q6_01	CETS1P54_03	EFC_Q6	RP58_01	ZF5_B
HNF4_01	GABP_B	CREBP1CJUN_01	VDR_Q3	HIC1_03
HNF4_DR1_Q3	ELK1_02	MIF1_01	CEBP_C	NRSF_01
PPAR_DR1_Q2	ELK1_01	VJUN_01	GCM_Q2	TFIII_Q6
COUP_01	NRF2_01	ATF3_Q6	TST1_01	HIC1_02
DR1_Q3	CETS168_Q6	ATF1_Q6	FOXJ2_02	AP2_Q6
HNF4_01_B	ETS1_B	T3R_01	NKX61_HAND1	AP2_Q3

Most significant tissue-matrix associations obtained when computing ⟨N⟩ for 200 bp proximal promoters and using GNF data for defining the tissue categories. Tissue-TF pairs with strong support in the literature are indicated in red while associations with likely function in the tissue are indicated in bold. For most tissues the same PFMs are recovered as when using EST derived tissues. Shown exceptions are thyroid and pituitary gland for which the corresponding PFMs could not be found.

MEF2. For PTF1 from a total of 14 pancreas specific target genes only four are shared between the two data sets while five targets are unique to each of the pancreas category derived from GNF and EST data. For MEF2 from a total of 24 skeletal muscle specific targets only eight are shared between the EST and GNF muscle categories. In general I observe that only a fraction of the designated targets of a given TF are assigned to the same tissue category in both the EST and GNF data set. It is encouraging to see that despite this small

overlap the sequence signals and the enrichment of target genes in either data set are strong enough to allow PASTAA to recover many of the functional tissue-TF associations. These findings also support the notion that GNF and EST data often rather complement each other.

6.4.6 Comparison to oPOSSUM, PAP and Clover

The results obtained by PASTAA were compared to the outcome of the three alternative methods, Clover (Frith et al., 2004), PAP (Chang et al., 2007) and oPOSSUM (Sui et al., 2005), each of which has been described in detail in Section 3.2. The oPOSSUM database contains pre-computed binding site hits only for the JASPAR database matrices (Sandelin et al., 2004) and for predefined promoter regions. Therefore, the 5-FP method (Rahmann et al., 2003) was utilized for the annotation of TF hits and subsequently the statistics introduced by oPOSSUM to detect any binding site enrichment in a given tissue category was applied (alternatively the balanced cutoff method was used to annotate binding site hits but the 5-FP method yielded better results). In order to make the results optimally comparable between PASTAA, Clover, PAP and oPOSSUM the tissue sets were restricted to only those genes with EST expression p-value < 10^{-6} and whose IDs could be unambiguously matched to entries in the PAP database. Usually less than 10% of the Ensembl IDs in a given EST derived tissue category could not be matched via either gene symbol or Refseq ID to a PAP entry. Both Clover and oPOSSUM were used with 200 bp proximal promoters as input while the PAP interface automatically uses larger promoter regions refined by phylogenetic footprinting.

The results of the comparative analysis are shown in Table 6.13 for five tissues. PAP detects well characterized associations for the liver and leukocyte specific tissue categories while for muscle and heart only two PFMs corresponding to muscle specific factors (MTATA and TEF) are detected. Clover tends to detect GC rich motifs (most notably SP1 and MAZ with consensus sequences GGGGCGGGG and GGGGAGGG, respectively) as highly enriched in all tested categories. The resulting ranking for more tissue specific factors suffers from this bias. oPOSSUM recovers many of the known muscle and liver specific associations and also detects several NFKB matrices for the leukocyte category among the top ten. However, also in the case of oPOSSUM many broadly acting factors with GC rich motifs including SP1 and UF1-H3beta (consensus motif GGTGGGGGAGGGGC) are detected as top ranking in several categories. This trend becomes stronger when extending the promoter regions to 500 bps (data not shown). Surprisingly, none of the three retina specific factors CRX, CHX10 and NRL was listed among the top ten matrices for the retina category by any of the alternative methods. However, Clover listed CRX and NRL among the top 20 matrices with position 11 and 13, respectively. In addition to the tissues shown in Table 6.13 PASTAA finds more

Table 6.13 – Comparison between Clover, oPOSSUM, PAP and PASTAA

	R	CLOVER	oPOSSUM	PAP	PASTAA
Striated muscle	1	SP1_Q2_01	**SRF_01**	TATA_01	**SRF_Q5_01**
	2	MAZ_Q6	**SRF_C**	T3R_Q6	**SRF_01**
	3	**MEF2_Q6_01**	**SRF_Q5_02**	**MTATA_B**	**SRF_Q5_02**
	4	TATA_01	**SRF_Q6**	SF1_Q6	**SRF_C**
	5	TBP_01	**SRF_Q4**	SPZ1_01	**MTATA_B**
	6	TBP_Q6	**MTATA_B**	PAX4_03	**MEF2_02**
	7	**MTATA_B**	TATA_01	MZF1_02	**SRF_Q6**
	8	WT1_Q6	PBX1_02	MZF1_01	HNF4_Q6_03
	9	**SRF_Q5_02**	PBX_Q3	HNF4_Q6_03	**SRF_Q4**
	10	SP1_01	**MEF2_02**	E2A_Q2	PR_02
Heart	1	SP1_Q4_01	**SRF_01**	SF1_Q6	**MEF2_Q6_01**
	2	SP1_Q2_01	**MEF2_02**	ERR1_Q2	**SRF_C**
	3	SP1_Q6	SP1_Q4_01	ER_Q6_02	**RSRFC4_01**
	4	GC_01	UF1H3BETA_Q6	T3R_Q6	**MTATA_B**
	5	SP1_Q6_01	**SRF_Q5_02**	TATA_01	**MEF2_02**
	6	SP1_01	SP1_Q6	HNF4_01	ZBRK1_01
	7	MAZ_Q6	GC_01	**TEF1_Q6**	**MEF2_03**
	8	WT1_Q6	**RSRFC4_01**	SP3_Q3	**SRF_01**
	9	KROX_Q6	SP1_Q6_01	HNF4_Q6_03	**RSRFC4_Q2**
	10	**MEF2_Q6_01**	ROAZ_01	XPF1_Q6	GR_Q6
Liver	1	SP1_Q4_01	**HNF4_Q6_01**	**CEBP_Q2_01**	**HNF4_Q6_01**
	2	SP1_Q2_01	**HNF1_01**	PBX1_03	**HNF1_01**
	3	GC_01	**HNF4_01**	**(C/EBP)**	**HNF4_01**
	4	SP1_Q6_01	**HNF1_Q6**	GR_Q6_01	**HNF1_Q6**
	5	SP1_Q6	**HNF1_Q6_01**	**HNF1_Q6**	**HNF1_C**
	6	CACBINDING_Q6	**HNF4_01_B**	**CEBPA_01**	**HNF4_01_B**
	7	MAZ_Q6	**DR1_Q3**	**CEBPB_02**	**HNF1_Q6_01**
	8	SP1_01	**HNF4_DR1_Q3**	**HNF4_Q6**	**HNF4_DR1_Q3**
	9	**HNF4_Q6_01**	**HNF1_C**	**HNF4_Q6_02**	**COUP_01**
	10	**DR1_Q3**	**COUP_01**	**HNF4_01**	**PPAR_DR1_Q2**
Retina	1	SP1_Q2_01	UF1H3BETA_Q6	SREBP1_Q6	GATA1_03
	2	CACBINDING_Q6	SP1_Q4_01	LFA1_Q6	**CRX_Q4**
	3	SP1_Q6_01	SP1_Q2_01	ZIC2_01	VMAF_01
	4	WT1_Q6	KROX_Q6	TFIII_Q6	SREBP1_02
	5	SP1_01	SP1_Q6	PAX4_03	**CHX10_01**
	6	MZF1_01	ZNF219_01	ZIC1_01	SREBP_Q3
	7	TBP_Q6	SP1_Q6_01	GATA1_03	ZNF219_01
	8	MOVOB_01	GC_01	MAZ_Q6	**NRL_HAND**
	9	TATA_01	AP2_Q6	MZF1_01	ZF5_01
	10	CDXA_01	WT1_Q6	ZIC3_01	PPARA_02
Leukocyte	1	SP1_Q4_01	SP1_Q4_01	**ETS_Q6**	**NFKAPPAB65_01**
	2	SP1_Q6_01	SP1_Q6	**PEA3_Q6**	**NFKAPPAB_01**
	3	SP1_Q2_01	GC_01	**PU1_Q6**	**NFKB_Q6_01**
	4	GC_01	SP1_Q6_01	**ETS_Q4**	**CREL_01**
	5	SP1_Q6	SP1_Q2_01	**(cREL)**	**ETS_Q6**
	6	**ETS_Q6**	SP1_01	**STAT6_02**	**NFKB_Q6**
	7	**STAT6_02**	**NFKB_Q6_01**	**(ETS)**	**NFKB_C**
	8	**PU1_Q6**	**NFKAPPAB65_01**	**CETS168_Q6**	**ELF1_Q6**
	9	CACBINDING_Q6	**NFKB_Q6**	**(ETS)**	SP1_Q6
	10	**ETS_Q4**	CACD_01	**CREL_01**	**ETS_Q4**

Top ranking PFMs according to PASTAA and three alternative approaches. Predictions corresponding to experimentally well characterized TF-tissue associations are shown in red. SP1 matrices are indicated in yellow while JASPAR matrices (used by PAP in addition to TRANSFAC) are shown in brackets. Results of Clover are first rank by p-values and in case of ties also on raw scores. **The PFM for NRL is not contained in TRANSFAC and JASPAR and was therefore not available for the PAP analysis.*

experimentally known associations also for the other compared tissue categories. For instance, for pituitary gland, PASTAA ranks the pituitary specific factor PIT1 as the top matrix while the other methods rank this factor only around position 10 and for pancreas PASTAA recovers PTF1 (pancreas specific transcription factor 1) as the top matrix, which is not listed among the top 10 matrices by any of the alternative approaches.

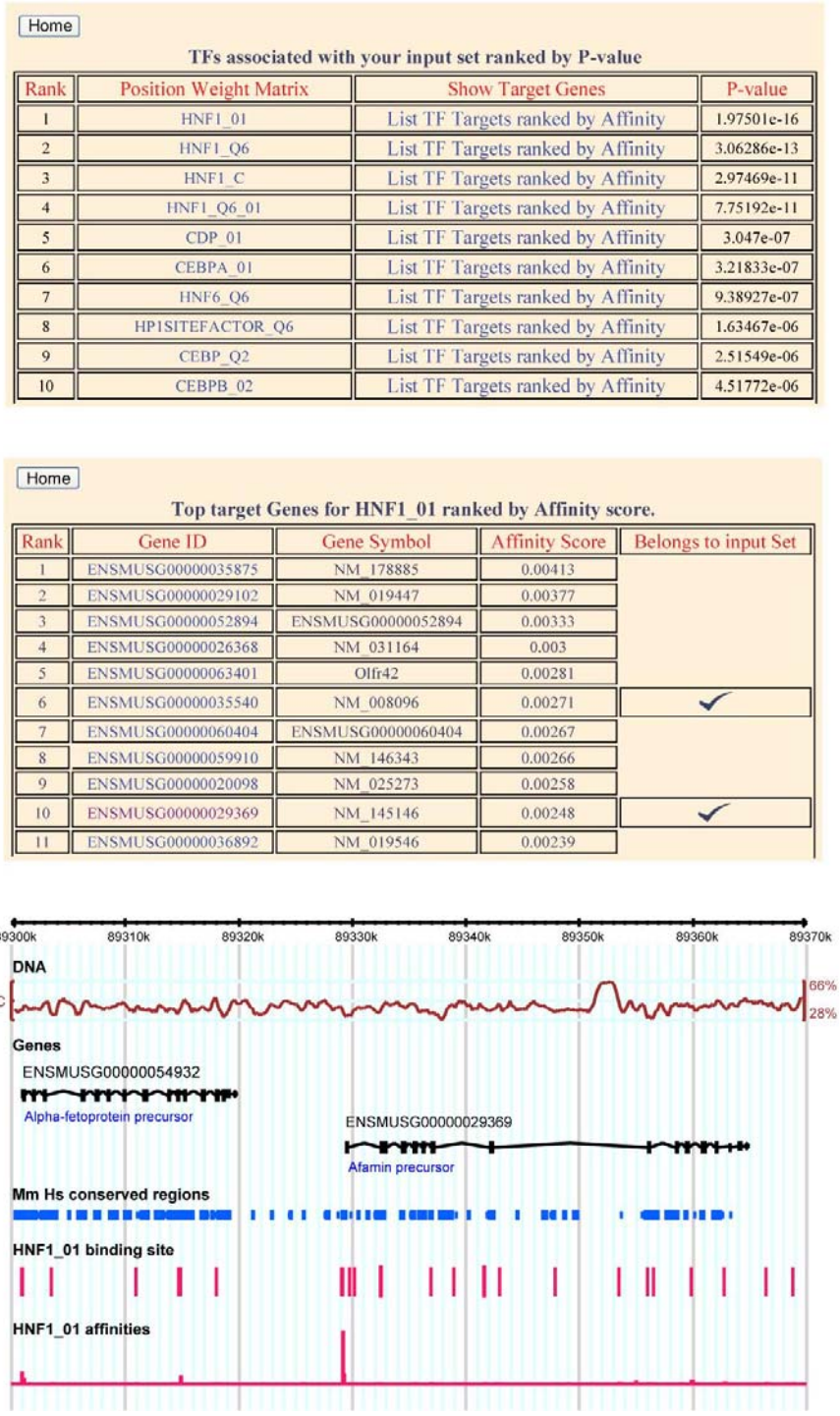
One might expect that PASTAA detects more of the experimentally known TF-tissue associations because the above analysis was carried out on the optimized 200 bp proximal promoters. However, this is not the case. In fact, PASTAA performed even more favourably if the sequence space was enlarged. When using 1kb long promoters without applying phylogenetic footprinting PASTAA retained many of the known TF-tissue associations for the above tissues while both oPOSSUM and Clover did not detect any of the experimentally known associations as top ranking.

In addition, oPOSSUM and Clover were applied to the HNF ChIP-chip and cMYC ChIP-PET data sets (see Section 6.3.3). For the HNF data sets both Clover and oPOSSUM produced similar rankings to those of PASTAA. In contrast, for the MYC data set Clover ranked the first MYC matrix at position 48 while all but one other MYC matrix were considered anti-correlated with the input set. Similarly, oPOSSUM ranked the first MYC matrix at position 28 while top ranking matrices correspond to immune related and heat shock factors. As a final test PASTAA was applied also to the muscle specific gene set used in the Clover publication and the muscle and NFKB microarray data sets used in the original oPOSSUM paper. For these data sets PASTAA obtained very similar results to those found by the corresponding methods. Over all the tested data sets PASTAA thus performed with greater specificity and sensitivity than the assessed alternative approaches.

6.5 Web implementation of PASTAA

With the help of Sean O'Keeffe I have implemented a simple and user-friendly website located at http://trap.molgen.mpg.de that allows users to find the TFs most strongly associated with their gene sets. The webpage thereby uses the hypergeometric test statistic together with the resampling procedure to compute accurate association p-values. The list of gene identifiers provided by the user can be composed of Ensembl Ids, Refseq Ids and gene symbols. For fast computation, TF affinities have been pre-computed for various promoter sizes for all human and mouse genes (Ensembl version 31). All genes not belonging to the input set are used as default background set.

Figure 6.13 – Web implementation for PASTAA



Home

TFs associated with your input set ranked by P-value

Rank	Position Weight Matrix	Show Target Genes	P-value
1	HNF1_01	List TF Targets ranked by Affinity	1.97501e-16
2	HNF1_Q6	List TF Targets ranked by Affinity	3.06286e-13
3	HNF1_C	List TF Targets ranked by Affinity	2.97469e-11
4	HNF1_Q6_01	List TF Targets ranked by Affinity	7.75192e-11
5	CDP_01	List TF Targets ranked by Affinity	3.047e-07
6	CEBPA_01	List TF Targets ranked by Affinity	3.21833e-07
7	HNF6_Q6	List TF Targets ranked by Affinity	9.38927e-07
8	HP1SITEFACTOR_Q6	List TF Targets ranked by Affinity	1.63467e-06
9	CEBP_Q2	List TF Targets ranked by Affinity	2.51549e-06
10	CEBPB_02	List TF Targets ranked by Affinity	4.51772e-06

Home

Top target Genes for HNF1_01 ranked by Affinity score.

Rank	Gene ID	Gene Symbol	Affinity Score	Belongs to input Set
1	ENSMUSG00000035875	NM_178885	0.00413	
2	ENSMUSG00000029102	NM_019447	0.00377	
3	ENSMUSG00000052894	ENSMUSG00000052894	0.00333	
4	ENSMUSG00000026368	NM_031164	0.003	
5	ENSMUSG00000063401	Olfr42	0.00281	
6	ENSMUSG00000035540	NM_008096	0.00271	✓
7	ENSMUSG00000060404	ENSMUSG00000060404	0.00267	
8	ENSMUSG00000059910	NM_146343	0.00266	
9	ENSMUSG00000020098	NM_025273	0.00258	
10	ENSMUSG00000029369	NM_145146	0.00248	✓
11	ENSMUSG00000036892	NM_019546	0.00239	

a) Primary result pages of the PASTAA web interface shown for an example input set of the 50 liver specific genes. After providing a list of gene identifiers PASTAA ranks all 593 PFMs from TRANSFAC according to their association with the input set. b) For a given PFM of interest a second page displays the ranking of all mouse or human genes according to predicted affinities. Finally, a link to a GBrowse implementation is provided, which displays the gene of interest and the annotated TF affinities by default.

The primary results page shows the ranking of all 593 PFMs according to the significance of their association with the input gene set (Figure 6.13a). The ranking of all mouse or human genes based on predicted affinities for a given TF can be obtained (Figure 6.13b) by clicking on the link next to the respective PFM. Genes in the ranked list belonging to the input set are hereby highlighted with a tick mark. The target genes are linked to a Genome Browser implemented by Stefan Haas and Sean O'Keeffe that automatically displays the gene structure, conservation between human and mouse as well as the annotated affinities for the chosen TF. Additional tracks displaying for instance TF binding site according to the balanced cutoff method (Rahmann et al., 2003) can be activated optionally (Figure 6.13c). For *de novo* annotation of sequences and matrices not provided on the webpage a stand alone C program is provided that requires as input a set of matrices, a FASTA file with sequences corresponding to a set of input genes and a FASTA file corresponding to a background set.

6.6 Discussion

In order to detect TFs regulating groups of genes I have embedded the TRAP model into a statistical framework called PASTAA that uses either a z-score statistic or a series of hypergeometric tests to assess the significance of the association between a given TF and a gene set. Although both statistical tests were highly successful in detecting TFs associated with a given ChIP-chip data set in yeast, the hypergeometric test has the advantage of being insensitive to outliers in the affinity predictions and to require only a minimum of *a priori* knowledge about gene set construction. In this setting PASTAA yields robust results against changes in the size of scanned promoter regions particularly when restricting the sequence space to evolutionary conserved blocks. Additionally the predicted associations are insensitive against considerable changes in the TRAP parameters. This stays in stark contrast to classical annotation methods where the number of predicted TF binding site hits and consequently the predicted associations often depend strongly on the chosen score threshold. Combining the hypergeometric test statistic with an iterative search for the optimal cutoffs applied to both input and target gene sets not only minimizes the amount of required *a priori* knowledge but further increases the robustness of the approach. In fact, using this method PASTAA was able to recover a majority of detected and verified TF-tissue associations even when using different expression data sets.

However, in many cases input genes will not be derived from ranked lists but from categorical data, such as all genes belonging to a given metabolic pathway. In an extreme case such a group may consist of only a single gene. Nevertheless, as was demonstrated for

several SRF target genes and promoters of direct autoregulatory factors PASTAA is highly successful in recovering corresponding regulators for individual genes. As an example for a larger gene group stemming from categorical data, PASTAA was supplied for instance with all genes assigned to the “early onset of diabetes” pathway in the KEGG database (Aoki et al., 2005) and recovered among the top 5 matrices the PFM representing the pancreas specific factor IPF1, which is a direct regulator of insulin expression (data not shown).

When applied to ranked lists of tissue specific genes PASTAA detects a comprehensive number of experimentally known TF-tissue pairs. These include not only associations such as SRF-heart, MEF2-muscle, HNF1-liver and NFKB-leukocyte, which are largely recovered also by the alternative computational approaches but also experimentally verified associations such as CRX-retina or PTF1-pancreas, which are not detected by Clover, PAP and oPOSSUM. Interestingly, meaningful associations are found by PASTAA not only for highly informative matrices such as NRSF_01 (Chen et al., 1998) with an information content of over 21 bits but also for matrices such as XPF1_Q6 (alternative PFM for PTF1) with an information content of only ≈ 3 bits.

Several lines of evidence support the identified TF-tissue associations. A) In accordance with previous studies (Xie et al., 2005; Tabach et al., 2007), which showed a general accumulation of regulatory signals near the transcription start site, PASTAA finds the region from 0 to 200 bps upstream of the TSS to yield most significant TF-tissue associations. This finding is further confirmed by the preferential location of windows with maximal affinity near the TSS. B) The verified associations obtained from the 200CPs are often confirmed by the rankings obtained when analyzing adjacent, non-overlapping promoter regions in particular from 200-400 bps upstream of the TSS. C) TFs are significantly more often over-expressed in their top ranking tissues than what would be expected based on results derived from random sequence sets. D) With a few exceptions, the majority of verified associations are found to be conserved between mouse and human. E) EST and GNF data yield identical top ranking TFs for the majority of tissue categories.

Aside from this evidence there is a considerable fraction of top ranking TF-tissue pairs well confirmed by literature. It should be stressed that only the most strongly supported TF-tissue associations (via literature or specific expression of the TF in the corresponding tissue) were considered for the analysis of expression preference and binding signal location in order to avoid any contamination of the set with false positives. However, there are some hundred additional top ranking associations that appear to be meaningful based on experimental findings. For instance, the hormone receptors GR (glucocorticoid receptor) and AR (androgen receptor) have testis as top associated tissue and are both considered to play

an important role in spermatogenesis (Weber et a. 2000, Collins et al 2003). Another example is the top association between cerebellum and NKX61, which is supported by literature (Qiu et al., 1998, Nelson et al., 2005) as well as expression of NKX61 in cerebellum.

In comparison to PASTAA and the other three tested methods there exist also several studies dedicated only to finding TFs regulating sets of tissue specific genes. For instance, by employing phylogenetic footprinting to select promoter regions and enhancer elements around gene loci, Pennacchio et al. (2007) were able to detect a number of experimentally verified associations for various tissues including liver, muscle and leukocytes. Alternatively, a study by Yu et al. (2006) investigated the tissue specificity of co-occurring binding sites for different pairs of TF. The authors were able to accurately assign several of their detected TF pairs to muscle, heart, liver, kidney, eye and lymph node. Finally, a group of works have focused on detecting binding signals within groups of tissue specific genes based on *de novo* motif finding (Smith et al. 2005; Huber et al. 2006). Such methods do not rely on known binding profiles but instead use motif finding programs such as MEME (Timothy et al., 2006) in order to detect overrepresented sequence patterns within promoters of co-expressed genes. Attempts to subsequently assign the found motifs back to PFMs of TFs with known function in the respective tissue were of limited success however. This is highlighted by a study of Huber et al. (2006) where combining four state of the art motif finding algorithms did not detect most of the experimentally confirmed TF-tissue associations including such well known cases as HNF1 and liver (Odom et al., 2004).

PASTAA compares well also against all of the above approaches by not only detecting the vast majority of experimentally verified TF-tissue associations recovered by these dedicated studies but also by recovering associations such as PTF1-pancreas not found by any of the alternative methods.

Despite the progress reported here for the detection of TF-tissue associations there are still a large number of tissues for which no known meaningful associations could be recovered. One reason for this might be the lack of EST expression data from several tissue or noise in the GNF data. In line with this notion, most verified and stable results were obtained for those tissues with large EST support. Given the dependency of the results on the quality and availability of expression data combining EST with GNF data in a sensible way (as suggested by Qian et al., 2005) might be of advantage. For instance, for tissues with low EST sampling depth one may form the union between the genes considered highly expressed in the tissue according to either microarray or EST measurements. On the other hand, in order to reduce noise in the data one could restrict the genes assigned to a given

tissue to only those with high level of support in both microarray and EST data. For instance, the most significant association between the muscle enhancer factor MEF2 and the muscle category is observed when restricting the input set to those genes, which are highly expressed in muscle according to both EST and microarray measurements (indicated by the intersection between EST and GNF data shown in Figure 6.12).

Apart from a lack of expression data, another reason for missing or perhaps downgraded associations could be the presence of sequence signals not directly linked to TF binding. For instance, many brain and testis specific genes have been found to possess promoters with high GC content. In turn computational methods tend to find a strong association between these tissues and PFMs representing GC rich binding motifs. However, such promoters might in fact be regulated primarily via methylation of CpG islands and changes in chromatin structure. Interestingly, this hypothesis can be partially substantiated by separating all genes into two groups, one possessing promoters with high CpG content and one with low CpG content. When using only the genes with low CpG promoters for the tissue analysis PASTAA detects nearly all of the presented TF-tissue associations. In contrast, when running the analysis on the high CpG promoters barely any experimentally known associations are recovered (data not shown). This indicates a profound distinction in the regulatory mechanisms controlling gene expression in different tissues and suggests that the discovery of regulating TFs might be facilitated by removing sequence signals that correspond to TF independent ways of regulation.

Finally, a last point to mention is that although the sequences up to 200 bps upstream of the TSS yielded the most significant results as well as many interpretable tissue-TF associations clearly a large number of functional high affinity sites will be located in enhancer regions far upstream or downstream of the respective TSS. For individual factors it might therefore prove advantageous to extend the search space to 5'-UTRs and intronic regions as was suggested for instance by Pennacchio et al. (2007). In this context it is interesting to mention that for the vast majority of tissue specific factors PASTAA does not detect significant binding signals downstream of the TSS (Dostie et al., 2006). In contrast, the ENCODE project (Birney et al., 2007) showed symmetrical distribution of TF binding sites around the TSS for a number of general transcription factors including E2F, MYC and the SWI/SNF chromatin-modifying complex. As was shown for NRSF (Figure 6.11) this might be the case also for some tissue specific factors. Therefore, it might prove advantageous to extend the PASTAA approach to not only adjust the size of the input and target gene set but also to automatically find the promoter region most enriched with tissue specific binding signals for a given TF.