Utilizing promoter pair orientations for HMM-based analysis of ChIP-chip data

Michael Seifert[1], Jens Keilwagen[1], Marc Strickert[1], and Ivo Grosse[2]

[1]Leibniz Institute of Plant Genetics and Crop Plant Research, Gatersleben, Germany
[2]Martin Luther University, Institute of Computer Science, Halle, Germany

seifert@ipk-gatersleben.de

Abstract: Array-based analysis of chromatin immunoprecipitation data (ChIP-chip) is a powerful technique for identifying DNA target regions of individual transcription factors. Here, we present three approaches, a standard log-fold-change analysis (LFC), a basic method based on a Hidden Markov Model (HMM), and an extension of the HMM approach to an HMM with scaled transition matrices (SHMM) to incorporate different promoter pair orientations. We compare the prediction of ABI3 target genes for the three methods and evaluate these genes using Genevestigator expression profiles and transient assays. We find that the application of the SHMM leads to a superior identification of ABI3 target genes. The software and the ChIP-chip data set used in our case study can be downloaded from http://dig.ipk-gatersleben.de/SHMMs/ChIPchip/ChIPchip.html.

1 Introduction

In recent years, array-based analysis of chromatin immunoprecipitation data (ChIP-chip) has become a powerful technique to identify DNA target regions of individual transcription factors. ChIP-chip was firstly applied to yeast by [RRW$^+$00] and [IHS$^+$01] based on promoter arrays. Nowadays, with the availability of sequenced genomes, ChIP-chip is mostly based on tiling arrays [JLG$^+$08]. The analysis of ChIP-chip data is challenging because of the huge data sets containing thousands of hybridization signals. Most of the available methods focus on the analysis of ChIP-chip tiling array data. Examples include a moving average method by [KvdLDC04], a Hidden Markov Model (HMM) approach by [LML05], or TileMap by [JW05] including both approaches.

Regarding *A. thaliana*, ChIP-chip is still far from being used routinely. In the trilateral project ARABIDOSEED, ChIP-chip based on promoter arrays was established for the seed-specific transcription factor ABI3. ABI3 is one of the fundamental regulators of seed development that is involved in controlling chlorophyll degradation, storage product accumulation, and desiccation tolerance [VCC05].

Here, we describe and compare three methods for the detection of transcription factor target genes from ChIP-chip data. The first method, which we abbreviate by LFC, is a standard log-fold change analysis in which the genes belonging to the promoters with the highest log-fold changes in the intersection of repeated experiments are considered to be putative target genes. The second method is based on a two-state (target promoter state and non-target promoter state) HMM. The principle architecture of the HMM follows the proposed two-state architecture by [LML05]. Our approach is extended in that way that all HMM parameters are directly learned from the ChIP-chip data. The HMM scores all promoters by the probability of being in the target promoter state, and we consider all genes belonging to promoters with the highest scores in the intersection of repeated experiments as putative target genes. The HMM allows statistical dependencies between ChIP-chip measurements of adjacent promoters along the chromosomes. The existence of such dependencies is clearly shown for ChIP-chip data of ABI3 in Fig. 1. We find that adjacent promoters in head-head orientation show significantly greater correlations than promoter pairs in head-tail, tail-head, or tail-tail orientation. The high correlations in ChIP-chip measurements of head-head promoter pairs can be explained by the array design: since proximal promoters but not complete intergenic regions are spotted. Thus, high positive correlations of measurements for head-head promoter pairs result from DNA segments of the intergenic region that bind to both promoter spots, or fragments of these segments where some of them bind to the one spot while the others bind to the other spot. The observation of correlations between ChIP-chip measurements of adjacent promoters motivates the extension of the HMM approach to an HMM with scaled transition matrices (SHMM). The general concept of SHMMs was developed by [Sei06] and applied to the analysis of tumor expression data by exploiting chromosomal distances of adjacent genes yielding to an improved detection of over-expressed and under-expressed genes. Here, we use this concept for discriminating head-head promoter pairs from other promoter pair orientations. The key assumption is that it is more likely for promoters in head-head orientation that both promoters are either target promoters or non-target promoters compared to other promoter orientations.

We use an ABI3 ChIP-chip data set for comparing the prediction of ABI3 target genes by the LFC, the HMM, and the SHMM method. We evaluate putative ABI3 target genes using (i) publicly available expression data from Genevestigator [ZHHHG04] and (ii) transient assays to test whether a putative target promoter is controlled by ABI3.

In general, good introductions to HMMs are given by [Rab89] or [DEKM98], extensions of standard HMMs to HMMs with transition matrices are described in [KSSW03], and some more details to SHMMs can be found in [Sei06]. A concept similar to SHMMs has been developed by [MD04] with an application to gene prediction.

2 Methods

2.1 Data acquisition and pre-processing

To determine target genes of the ABI3 transcription factor the ChIP-chip technique by [RRW$^+$00] and [IHS$^+$01] was applied to *A. thaliana* wildtype seeds. Isolated DNA fragments bound by ABI3 were amplified, radio-labeled, and hybridized to a macroarray con-

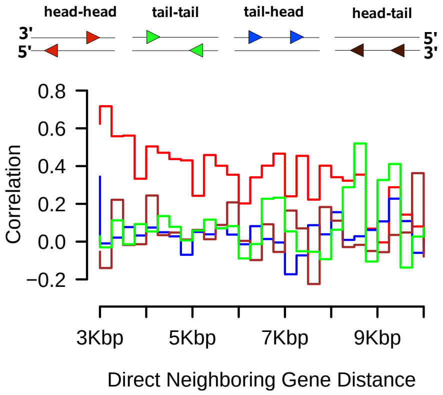

Figure 1: Pearson's correlations for the four promoter pair orientations based on log-ratios of ABI3 ChIP-chip experiments in steps of 250bp within the corresponding gene pair distance interval $[3, 10]$Kbp. A triangle represents a promoter and the orientation of its tip describes the reading direction of the gene belonging to this promoter.

taining 11,904 promoters of *A. thaliana*. The corresponding control sample was obtained from the input chromatin of the wildtype seeds by fragmentation, amplification, labeling, and hybridization to another promoter macroarray. In total, each of these two experiments was repeated five times. In a first normalization step, we center the median of each experiment to zero and perform a quantile normalization [BIAS03] separately for the ABI3 ChIP-chip experiments and the control experiments. In a second step, we combine each normalized ABI3 ChIP-chip experiment with its corresponding control experiment by calculating the log-ratio $o_t = I_{\mathit{ABI3}}(t) - I_{\mathit{Control}}(t)$ for all promoters t, where $I_{\mathit{ABI3}}(t)$ is the $\log_2$-signal intensity of promoter t in the ABI3 ChIP-chip experiment, and $I_{\mathit{Control}}(t)$ is the $\log_2$-signal intensity of promoter t in the control experiment. We map all of the log-ratios of such an experiment combination to their corresponding positions in the genome of *A. thaliana* based on the TAIR7 genome annotation, resulting in one ChIP-chip profile $o = o_1, \ldots, o_T$ per chromosome. As *A. thaliana* has five chromosomes 25 ChIP-chip profiles were obtained from the five replicates.

2.2 Standard Log-Fold-Change analysis (LFC) for target gene detection

The log-ratio of a promoter characterizes the potential of the gene belonging to this promoter to be a target gene of the ABI3 transcription factor. Thus, we expect that putative ABI3 target genes have log-ratios that are significantly greater than zero in repeated experiments. For each of the five replicated experiments, we create a list containing all of the promoter identifiers of the ChIP-chip profiles of the five chromosomes in decreasing order of their log-ratios. That is, promoters with log-ratios significantly greater than zero are at the top of this list. We use these five lists to determine the intersection of the top

k candidate promoters of each list. This proceeding allows to assess the degree of reproducibility between the five replicates. We interpret all genes belonging to the promoters in the intersection as putative target genes of ABI3.

2.3 Hidden Markov Model (HMM) for target gene detection

HMM description: We use a two-state HMM $\lambda = (S, \pi, A, E)$ with Gaussian emission densities for the genome-wide detection of putative ABI3 target genes. The basis of this HMM is the set of states $S = \{-, +\}$. State $-$ corresponds to a promoter that is not a target of ABI3, and state $+$ corresponds to a promoter that is a target of ABI3. We denote the state of promoter t by $q_t \in S$, and we assume that a state sequence $q = q_1, ..., q_T$ belonging to a ChIP-chip profile o is generated by a homogeneous Markov model of order 1 with start distribution $\pi = (\pi_-, \pi_+)$ and stochastic transition matrix $A = (a_{ij})_{i,j \in S}$ where $\pi_-, a_{--}, a_{++} \in (0, 1)$, $\pi_+ = 1 - \pi_-$, $a_{-+} = 1 - a_{--}$, and $a_{+-} = 1 - a_{++}$. The state sequence is assumed to be not observable, i.e. hidden, and the log-ratio o_t of promoter t is assumed to be drawn from a Gaussian emission density, whose mean and standard deviation depend on state q_t. We denote the vector of emission parameters by $E = (\mu_-, \mu_+, \sigma_-, \sigma_+)$ with means μ_- and μ_+, and standard deviations σ_- and σ_+ for the Gaussian emission density $b_i(o_t) = 1/(\sqrt{2\pi}\sigma_i) \exp(-0.5(o_t - \mu_i)^2/\sigma_i^2)$ of log-ratio o_t given state $i \in S$.

HMM initialization: In general, an initial HMM has to discriminate ABI3 target promoters from non-target promoters with respect to their log-ratios in the ChIP-chip profile. Hence, a histogram of log-ratios of all five replicates helps to find good initial HMM parameters. The choice of initial parameters addresses the presumptions that the proportion of non-target promoters is much higher than that of target promoters, and that the number of successive non-target promoters is also much higher than the number of successive target promoters. In our case study we use $\pi_- = 0.9$ resulting in an initial start distribution $\pi = (0.9, 0.1)$. Thus, we choose an initial transition matrix A with equilibrium distribution π. That is, we set $a_{--} = 1 - s/\pi_-$ and $a_{++} = 1 - s/\pi_+$ with respect to the scale parameter $s = 0.05$ to control the state durations. We characterize the states by proper means and standard deviations using initial emission parameters $\mu_- = 0$, $\mu_+ = 2$, $\sigma_- = 1$, and $\sigma_+ = 0.5$. We refer to the initial HMM by λ^1.

HMM training: We train the initial HMM based on all ChIP-chip profiles using a maximum a posteriori (MAP) variant of the standard Baum-Welch algorithm ([Rab89], [DEKM98]). This algorithm is part of the class of EM algorithms ([DLR77]), which iteratively maximize their optimization function. With respect to the underlying biological question, the choice of the prior influences the quality of the trained HMM. We include biological a priori knowledge into the MAP training using a Dirichlet prior with hyper-parameters $\vartheta_- = \vartheta_+ = 2$ for start distribution π, a product of Dirichlet priors with hyper-parameters $\vartheta_{ab} = 1$ with $a, b \in S$ for transition matrix A, and a product of Normal-

Gamma priors for emission parameters E with hyper-parameters $\eta_- = 0$ and $\eta_+ = 2$ (a priori means), $\epsilon_- = \epsilon_+ = 1,000$ (scale of a priori means), $r_- = 1$ and $r_+ = 100$ (shape of standard deviations), and $\alpha_- = \alpha_+ = 10^{-4}$ (scale of standard deviations). The choice of these prior parameters ensures a good characterization of both HMM states. On that basis we iteratively maximize the posterior of the HMM λ^h given all ChIP-chip profiles resulting in new HMM parameters λ^{h+1}. We stop the MAP training if the increase of the log-posterior of two successive MAP iterations is less than 10^{-9}.

HMM target gene detection: The state $+$ of the trained HMM λ models the potential of promoters to be targets of ABI3. To quantify this potential we calculate the probability $\gamma_t(+) = P[Q_t = +|O = o, \lambda]$ for each promoter t within a ChIP-chip profile o to be a target promoter. This state posterior of state $+$ is computed using the Forward-Backward procedures of HMMs ([Rab89], [DEKM98]). For each of the five replicated experiments we create a list containing all of the promoter identifiers of the ChIP-chip profiles of the five chromosomes in decreasing order of their state posteriors $\gamma_t(+)$. We use these five lists to determine the intersection of the top k candidate promoters of each list. In analog to the standard LFC approach, we interpret all genes belonging to the promoters in the intersection as putative target genes of ABI3.

2.4 Hidden Markov Model with scaled transition matrices (SHMM) for target gene detection

SHMM description: The general concept of SHMMs enables us to analyze ChIP-chip profiles in the context of orientations of neighboring genes on the DNA. Two directly neighboring genes on DNA occur either in head-head, tail-tail, tail-head, or head-tail orientation to each other. Among these orientations the head-head orientation is of prime importance for the analysis of promoter array data. In this orientation the two corresponding genes have the potential to share a common promoter region depending on the distance between these genes. This fact in combination with the observation that the log-ratios of promoters for genes in head-head orientation show significantly higher correlations compared to all other orientations is the basis to design a specific SHMM. We assume that it is more likely for two genes in head-head orientation to show the same promoter status, that means either ABI3 target or not, in comparison to all other orientations. For that reason we assign to each pair of successive promoters t and $t+1$ of a chromosome one promoter pair orientation class $c(d_t)$ depending on the orientation of both promoters to each other in combination with the chromosomal distance d_t of the two genes belonging to these promoters. The promoter pair orientation class of successive promoters t and $t+1$ is

$$c(d_t) = \begin{cases} 2, & t \text{ and } t+1 \text{ are head-head and } d_t \leq b \\ 1, & \text{otherwise} \end{cases}$$

using a pre-defined distance threshold $b \in \mathbb{N}$. We incorporate these information into a two-state SHMM $\lambda_L = (S, \pi, A, \vec{f}, E)$ with $L = 2$ promoter pair orientation classes to detect putative ABI3 target genes. The parameters S, π, A, and E are defined like in the HMM approach, and $\vec{f} = (f_1 := 1, f_2)$ with $f_2 \in \mathbb{R}^+$ and $f_2 > f_1$ is the vector of scaling factors. In contrast to the standard HMM approach, we now assume that the state sequence q of a ChIP-chip profile o is generated by an inhomogeneous Markov model of order 1 with start distribution π and two scaled stochastic transition matrices A_1 and A_2 for discriminating head-head orientations from others based on the promoter pair orientation classes. The transition matrix A_l with $l \in \{1, 2\}$ is defined by

$$A_l \quad = \quad \frac{1}{f_l} \begin{pmatrix} a_{--} - 1 + f_l & a_{-+} \\ a_{+-}, & a_{++} - 1 + f_l \end{pmatrix}$$

with respect to the scaling factor f_l that scales the expected state duration of state $i \in S$ in A from $1/(1-a_{ii})$ to $f_l/(1-a_{ii})$ in A_l. A transition from state q_t to state q_{t+1} is achieved by using the corresponding transition matrix $A_{c(d_t)}$ based on the integrated promoter pair orientation class $c(d_t)$. The self-transition probability of each state $i \in S$ increases strictly from matrix A_1 to A_2, and thus, for a head-head promoter pair that is modeled by A_2 it is more likely that both promoters are targets or no targets of ABI3 compared to other promoter pairs modeled by A_1. The log-ratios of promoters are modeled as described in the HMM approach.

SHMM initialization: The basic initialization of the SHMM is done like for the HMM. In addition to that, we must choose a distance threshold b for the promoter pair orientation classes and a scaling factor f_2 to specify the degree of differentiation between head-head orientation modeled by A_2 and all others modeled by A_1. Motivated by Fig. 1 we always use $b = 9$Kbp in our case study because in greater chromosomal distance the correlations of head-head promoter pairs do not significantly differ from others. Moreover, we consider all f_2 from 1.1 to 10 in steps of 0.1.

SHMM training: The SHMM is trained like the HMM using the MAP variant of the Baum-Welch algorithm with identical prior hyper-parameters. The only difference between both models occurs during the estimation of their transition matrices. Details of the parameter estimation are described by [Sei06].

SHMM target gene detection: The putative target genes of ABI3 are determined in analog to the HMM approach. The calculation of the state posterior $\gamma_t(+)$ is now done with respect to the promoter pair orientation classes using the Forward-Backward procedures of HMMs.

3 Results and discussion

3.1 Study of differences between HMM and SHMMs

The HMM approach enables us to analyze ChIP-chip data in the context of chromosomal locations of promoters, and the application of SHMMs extends this analysis by discriminating different types of promoter pair orientations. In a first study, we investigate how SHMMs behave compared to the standard HMM. For that reason, we use the Viterbi algorithm ([Rab89], [DEKM98]) to compare the most likely state sequence q for a ChIP-chip profile o under the trained HMM to that of all trained SHMMs with scaling factor f_2 increasing from 1.1 to 10 in steps of 0.1. Here, the annotation of a promoter t with log-ratio o_t is given by $q_t \in S$, which we interpret as this promoter is either a putative ABI3 target or not. The scaling factor allows to directly influence the annotation behavior for head-head promoters. That is, the higher f_2 the more likely it is that both promoters of such head-head pairs are either putative ABI3 targets or not, and the closer we choose f_2 to one the closer is the annotation behavior of the SHMM to that of the HMM. The results are illustrated in Fig. 2a. We observe that the number of head-head promoter pairs where both promoters of such a pair have identical annotations increases for increasing scaling factor f_2, and as consequence the number of head-head promoter pairs where both promoters of such a pair have different annotations decreases. Obviously, each change in the annotation of a head-head promoter pair leads either to a change in the annotation of the upstream, downstream, or both of these promoter pairs. We see that the number of non-head-head promoter pairs where both promoters of such a pair are annotated as putative ABI3 targets decreases only slightly for SHMMs with increasing scaling factor f_2 compared to the HMM. We clearly see substantially more decrease in the number of non-head-head promoter pairs where both promoters of such a pair are annotated as putative non-target promoters for SHMMs with increasing scaling factor f_2 in relation to the HMM. Consequently, the number of non-head-head promoter pairs where both promoters of such a pair have different annotations increases with increasing scaling factor f_2. This study demonstrated that the annotation results of SHMMs can differ significantly from that of the HMM resulting in a more general model for the prediction of putative target genes.

3.2 Comparison of LFC, HMM, and SHMM to predict ABI3 target promoters

We use the LFC method for scoring putative ABI3 target promoters based on the log-ratios of the promoters neglecting chromosomal locations and promoter pair orientations. For comparison, we make use of the HMM that models chromosomal locations of promoters and the SHMM that models chromosomal locations and orientations of promoter pairs whereas both methods score putative ABI3 target promoters via the state posterior of state $+$. In this comparison study we set the threshold for the maximal number of candidates in a top list to 200 because the mean log-ratio of 1.06 at this level is already relatively small, and beyond, at a threshold of 300 we did not get new putative ABI3 target genes by the three methods. Moreover, we use the SHMM with scaling factor $f_2 = 4$ in all further analyses because this model is already quite different from the standard HMM (Fig. 2a), and the comparison of this model to SHMMs with scaling factor $f_2 = 6$ and $f_2 = 10$ yielded identical target promoters. For each approach, we score all five experiments to determine the intersection of putative ABI3 target promoters for the top 50, 100, 150 and 200 candidates under these experiments. Then, we use Venn diagrams to directly compare the candidate promoters for these four top lists under all three methods. The results are shown in Fig. 2b. We observe that the SHMM predicted the greatest number of putative ABI3 target promoters, whereas the LFC method predicted the smallest number. When we consider the Venn diagrams from the top 100 list to the top 200 list all candidates that are predicted by the LFC method are also completely predicted by both the HMM and the SHMM. In addition to this, the candidates additionally predicted by the HMM from the top 150 list to the top 200 list are completely predicted by the SHMM. In summary, this emphasizes that the SHMM approach tends to be more general in the prediction of putative ABI3 target promoters than the HMM and the LFC approach.

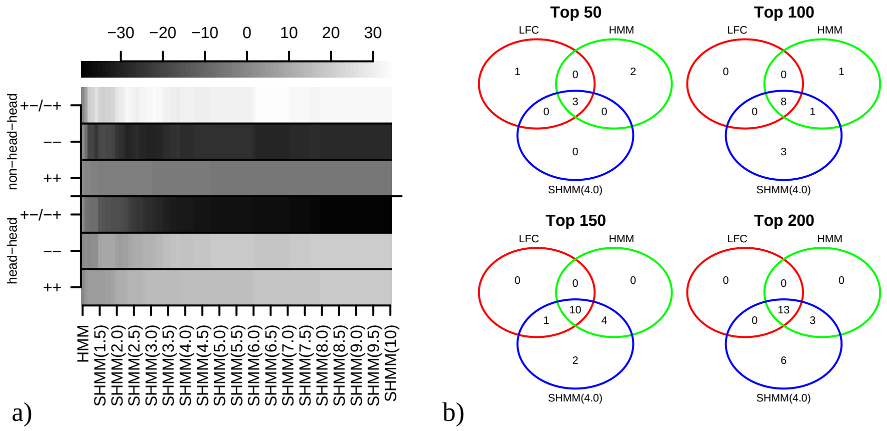

Figure 2: **a)**: Frequencies of promoter pair annotations of the trained SHMM(f_2) with scaling factor $f_2 \in [1.1, 10]$ in steps of 0.1 in relation to the trained HMM based on Viterbi annotations. The grey gradient in the upper part expresses the quantity of annotation differences whereas the HMM is encoded by the grey with value zero. The annotations $++$, $--$, and $+-/-+$ of promoter pairs mean that either both promoters are putative targets, non-targets, or only one promoter is a putative target of ABI3. **b)**: Venn diagrams to compare putative ABI3 target promoters predicted by the LFC method, the trained HMM, and the trained SHMM(4.0).

3.3 Gene expression analysis of putative ABI3 target genes belonging to predicted ABI3 target promoters

Next we investigate how putative target genes are regulated by ABI3. Therefore, we use Genevestigator [ZHHHG04] as independent source of *A. thaliana* gene expression data to analyze putative target genes. In Genevestigator, ABI3 is mainly expressed within the categories inflorescence, silique, and seed. Based on that, we quantify the expression of all putative target genes by dividing the sum of expression values within these three categories by the sum of expression values in all categories. This provides a quantitative measure, which we call Genevestigator score, for analyzing how a putative ABI3 target gene follows the expression profile of ABI3. Additionally, transient assays have been performed to test whether putative target promoters in fusion with the glucuronidase (GUS) reporter gene react on ABI3. The results are shown in Tab. 1. Calculating the Genevestigator score, 16 of 22 putative target genes show significantly high scores at the level of the 95%-quantile 0.15 based on the distribution of the Genevestigator scores for 1,000 randomly selected genes. The promoters of these 16 genes have been tested in transient assays, and we find that 15 of these promoters can activate the GUS expression through ABI3, and the promoter of gene T21 shows nearly a two-fold repression of the GUS expression. Interestingly, the genes T21 and T22 are in head-head orientation to each other, and so they have the potential to share a common promoter region. Based on the results of the transient assays the first gene might be repressed while the second is activated. Hence, it seems that activation and repression signals can be transmitted by ABI3 to these two target genes in head-head orientation via a potential common promoter region. Additionally, we point out that only the SHMM approach was able to predict 3 of these 15 target genes activated by ABI3 and the one target gene repressed by ABI3. In contrast to these 16 target genes, the 6 remaining putative target genes do not significantly differ in their Genevestigator scores at the level of the 5%-95%-quantile range $[0.02, 0.15]$ based on the distribution of the Genevestigator scores for the 1,000 randomly selected genes. Interestingly, 5 of these 6 putative target genes are in head-head orientation to one of the previous target genes activated by ABI3, and so the potential common promoter region can already receive signals from ABI3. Next we address the question if these 6 putative ABI3 target genes are also under control of ABI3 via the potential common promoter region. To test this hypothesis, the promoters of 4 of these 6 putative target genes have been tested in transient assays. The promoters of the genes T2 and T11 show a low activation of the GUS expression, the promoter of gene T13 shows a two-fold repression of the GUS expression, and the promoter of gene T9 does not seem to react on ABI3. In addition to this, gene T13 is in head-head orientation with gene T23 that is not represented by its own proximal promoter fragment on the promoter arrays. The Genevestigator score of T23 is significantly higher than those of the 1,000 random genes at the level of the 95%-quantile, and the promoter of this gene shows activation of the GUS expression in a transient assay. Hence, this gene pair seems to behave like the gene pair T21 and T22. In summary, independent gene expression profiles from Genevestigator give first hints which genes might be activated by ABI3. Additionally, transient assays help to validate this results if the underlying test system is capable of simulating the natural situation in seeds. Twenty percent of the ABI3 activated target genes with high Genevestigator scores could only be predicted through the application of the SHMM approach and would have been missed using the LFC or HMM approach. Moreover, the SHMM predicted over forty percent more putative ABI3 target genes compared to the LFC method. For these 9 genes the promoters of 7 have been tested in transient assays whereas 1 promoter does not react, 1 represses the GUS expression, and the 5 others activate the GUS expression. This results emphasize the relevance of SHMMs in the detection of ABI3 target genes.

ID	LFC	HMM	SHMM(4.0)	Genevestigator	Transient Assay
T1	1	1	1	0.94	5
T2	1	1	1	0.11	2.5
T3	1	1	1	0.86	12
T4	0	0	1	0.03	-
T5	0	0	1	0.39	3
T6	1	1	1	0.72	15
T7	1	1	1	0.90	7
T8	0	0	1	0.46	12
T9	0	0	1	0.07	1
T10	0	0	1	0.95	6
T11	0	1	1	0.09	2
T12	1	1	1	0.74	24
T13	1	1	1	0.09	0.4
T14	1	1	1	0.93	8
T15	0	1	1	0.10	-
T16	1	1	1	0.95	27
T17	1	1	1	0.98	27
T18	0	1	1	0.98	27
T19	1	1	1	0.98	27
T20	1	1	1	0.57	8
T21	0	0	1	0.20	0.6
T22	1	1	1	0.81	30

Table 1: Overview of predicted ABI3 target genes at the level of the top 200 candidates in Fig. 2b. The ID column contains anonymized target gene identifiers (our biologists prepare a manuscript discussing target genes). The numbers 1 and 0 in the method columns LFC, HMM, and SHMM(4.0) encode whether a gene is predicted or missed. Genevestigator quantifies the gene expression of a target gene within the categories inflorescence, silique, and seed as described in Section 3.3. Transient Assay contains the measured fold-change for a target gene promoter under ABI3 expression vs. target gene promoter lacking ABI3 expression.

4 Conclusions and outlook

We introduced the LFC, the HMM, and the SHMM approach for the analysis of ChIP-chip promoter array data and compared these methods on ABI3 ChIP-chip data. The motivation for the usage of HMMs is based on the observation of positive correlations between ChIP-chip measurements of adjacent promoters on the DNA (Fig. 1). Especially, the SHMM approach is motivated by the fact that ChIP-chip measurements of head-head promoter pairs show significantly higher correlations than those of other promoter pair orientations. Based on SHMMs, we demonstrated that discriminating promoters in head-head orientations from other promoter orientations can lead to significantly different predictions of target and non-target promoters compared to the HMM (Fig. 2a). Regarding all three methods, the SHMM predicted the highest number of putative ABI3 target promoters and all target promoters predicted by the LFC or the HMM have been included (Fig. 2b), but the number of predicted putative ABI3 target promoters is not an optimal criterion to decide which of the methods should be preferred. For this reason, we used publicly available expression profiles from Genevestigator to analyze how a putative target gene follows the expression profile of ABI3, and transient assays have been performed to test whether the promoter of a putative target gene reacts on ABI3 (Tab. 1). We showed that expression data from Genevestigator can give first hints which genes might be activated by ABI3, and that the validation can be done by transient assays. Twenty percent of the target genes with significantly high Genevestigator scores and activation in transient assays could only be predicted by the SHMM and would have been missed by the LFC or HMM approach. In total, the SHMM predicted more than forty percent more putative target promoters (9 of 22) compared to the LFC method. Seven of these promoters have been tested in transient assays whereas one promoter does not react, one represses the GUS expression, and the five others activate the GUS expression. Taking this together, we conclude that the SHMM can be seen as the more general model that should be preferred for the prediction of ABI3 target genes. We conjecture that the proposed SHMM might possibly be useful for the analysis of other promoter array ChIP-chip data.

In the future, the study of seed development continues. For instance, we are awaiting ChIP-chip data of the transcription factors LEC1, LEC2, and FUS3. This will provide us first insights into the transcriptional regulatory network involved in seed development. In cooperation with us, our biologists prepare a manuscript with details to the ABI3 ChIP-chip experiments including the discussion of ABI3 target genes.

5 Acknowledgments

We thank the groups of Lothar Altschmied, Helmut Bäumlein, and Udo Conrad and especially Urs Hähnel and Gudrun Mönke for ChIP-chip data, transient assays, and valuable discussions. This work was supported by the BMBF grants 0312706A and 0313155, and by the Ministry of culture Saxony-Anhalt grant XP3624HP/0606T.

References

[BIAS03] BM Bolstad, RA Irizarry, M Astrand, and TP Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, 2003.

[DEKM98] R Durbin, S Eddy, A Krogh, and G Mitchision. *Biological sequence analysis - Probabilistic models of proteins and nucleic acids*. Cambridge University Press, 1998.

[DLR77] A Dempster, N Laird, and D Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[IHS$^+$01] VR Iyer, CE Horak, CS Scafe, D Botsein, M Snyder, and PO Brown. Genomic binding sites of the yeast cell-cycle transcription factors SFB and MBF. *Nature*, 409:533–538, 2001.

[JLG$^+$08] DS Johnson, W Li, DB Gordon, A Bhattacharjee, B Curry, and L Brizuela et al. Systematic evaluation of variability in ChIP-chip experiments using predefined DNA targets. *Genome Res*, 18:393–403, 2008.

[JW05] H Ji and WH Wong. TileMap: create chromosomal map of tiling array hybridizations. *Bioinformatics*, 21(18):3629–3636, 2005.

[KSSW03] B Knab, A Schliep, B Steckemetz, and B Wichern. Model-based clustering with Hidden Markov Models and its application to financial time-series data. *In M. Schader, W. Gaul, and M. Vichi, editors, Between Data Science and Applied Data Analysis, Springer*, pages 561–569, 2003.

[KvdLDC04] S Keles, MJ van der Laan, S Dudoit, and SE Cawley. Multiple testing methods for ChIP-chip high density oligonucleotide array data. *Working Paper Series 147*, 2004. U.C. Berkeley Division of Biostatistics, University of California, Berkeley, CA.

[LML05] W Li, CA Meyer, and XS Liu. A hidden Markov model for analyzing ChIP-chip experiments on genome tiling arrays and its application to p53 binding sequences. *Bioinformatics*, 21:i274–i282, 2005.

[MD04] I M Meyer and R Durbin. Gene structure conservation aids similarity based gene prediction. *Nucleic Acids Research*, 32(2):776–783, 2004.

[Rab89] L Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[RRW$^+$00] B Ren, F Robert, JJ Wyrick, O Aparicio, EG Jennings, I Simon, J Zeitlinger, J Schreiber, N Hannett, E Kanin, TL Volkert, CJ Wilson, SP Bell, and RA Young. Genome-Wide Location and Function of DNA Binding Proteins. *Science*, 290(5500):2306–2309, 2000.

[Sei06] M Seifert. Analysing Microarray Data Using Homogeneous And Inhomogeneous Hidden Markov Models. Diploma Thesis; Martin Luther University; seifert@ipk-gatersleben.de, 2006.

[VCC05] J Vicente-Carbajosa and P Carbonero. Seed maturation: developing an intrusive phase to accomplish a quiescent state. *Int. J. Dev. Biol.*, 49:645–651, 2005.

[ZHHHG04] P Zimmerman, M Hirsch-Hoffman, L Hennig, and W Gruissem. GENEVESTIGATOR. Arabidopsis Microarray Database and Analysis Toolbox. *Plant Physiol.*, 136:2621–2632, 2004.