BIOINFORMATICS ORIGINAL PAPER
Vol. 25 no. 16 2009, pages 2118–2125
doi:10.1093/bioinformatics/btp276

Gene expression

Utilizing gene pair orientations for HMM-based analysis of promoter array 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, Data Inspection Group, Corrensstrasse 3, 06466 Gatersleben, and [2]Martin Luther University, Institute of Computer Science, Von-Seckendorff-Platz 1, 06120 Halle, Germany

Received on October 29, 2009; Revised on February 27, 2009; Accepted on April 21, 2009

Advance Access publication April 28, 2009

Associate Editor: Trey Ideker

ABSTRACT

Motivation: Array-based analysis of chromatin immunoprecipitation (ChIP-chip) data is a powerful technique for identifying DNA target regions of individual transcription factors. The identification of these target regions from comprehensive promoter array ChIP-chip data is challenging. Here, three approaches for the identification of transcription factor target genes from promoter array ChIP-chip data are presented. We compare (i) a standard log-fold-change analysis (LFC); (ii) a basic method based on a Hidden Markov Model (HMM); and (iii) a new extension of the HMM approach to an HMM with scaled transition matrices (SHMM) that incorporates information about the relative orientation of adjacent gene pairs on DNA.

Results: All three methods are applied to different promoter array ChIP-chip datasets of the yeast *Saccharomyces cerevisiae* and the important model plant *Arabidopsis thaliana* to compare the prediction of transcription factor target genes. In the context of the yeast cell cycle, common target genes bound by the transcription factors ACE2 and SWI5, and ACE2 and FKH2 are identified and evaluated using the Saccharomyces Genome Database. Regarding *A.thaliana*, target genes of the seed-specific transcription factor ABI3 are predicted and evaluate based on publicly available gene expression profiles and transient assays performed in the wet laboratory experiments. The application of the novel SHMM to these two different promoter array ChIP-chip datasets leads to an improved identification of transcription factor target genes in comparison to the two standard approaches LFC and HMM.

Availability: The software of LFC, HMM and SHMM, the ABI3 ChIP–chip dataset, and Supplementary Material can be downloaded from http://dig.ipk-gatersleben.de/SHMMs/ChIPchip/ChIPchip.html.

Contact: seifert@ipk-gatersleben.de

1 INTRODUCTION

In recent years, array-based analysis of chromatin immunoprecipitation (ChIP-chip) data has become a powerful technique to identify DNA target regions of individual transcription factors. ChIP-chip was first applied to yeast by Ren *et al.* (2000) and Iyer *et al.* (2001) based on promoter arrays. Nowadays, with the availability of sequenced genomes, ChIP-chip is predominantly based on tiling arrays (Johnson *et al.*, 2008). The analysis of ChIP-chip data is challenging, because of the huge datasets containing thousands of hybridization signals. Most available methods focus on the analysis of tiling array ChIP-chip data to predict the DNA-binding targets of DNA-binding proteins like transcription factors or histones. Examples include a moving average method by Keles *et al.* (2004), a Hidden Markov Model (HMM) approach by Li *et al.* (2005), TileMap by Ji and Wong (2005) using moving averages or an HMM to account for information of adjacent probes, or PMT by Chung *et al.* (2007) that integrates a physical model to correct for probe-specific behavior. Recently, a new HMM approach was developed by Humburg *et al.* (2008), outperforming TileMap in the

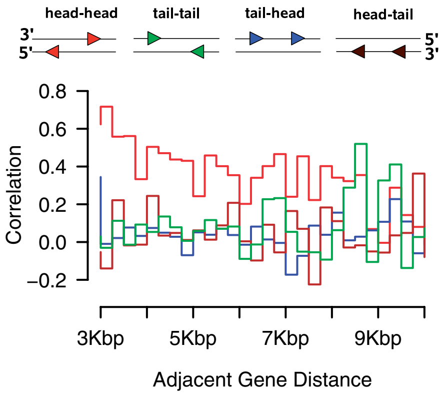

Fig. 1. Pearson correlations of promoter array ChIP-chip measurements of the transcription factor ABI3 in the context of the four gene pair orientations head–head, tail–tail, tail–head, and head–tail of two directly adjacent genes on DNA in distances of 3 Kbp up to 10 Kbp in steps of 250 bp. Genes are represented by triangles, and the orientation of the tip of a triangle defines the reading direction of a gene. The promoter fragment of a gene in the ABI3 dataset is always located in 3' direction of the gene. The ChIP-chip measurement of a gene is the $\log_2$-ratio of immunoprecipitated DNA for ABI3 to input control DNA that is measured for the corresponding promoter of the gene. The intergenic region between two genes in head–head orientation is represented by two promoter fragments, one for each gene. Depending on the distance between these two genes the extracted DNA segments in the immunoprecipitated sample and in the input DNA sample can bind to both promoter fragments of these two head–head genes leading to significantly higher correlations for genes in head–head orientation in comparison to all other gene pair orientations.

[*]To whom correspondence should be addressed.

Table 1. Pearson correlations of promoter array ChIP-chip measurements of transcription factors ACE2, SWI5 and FKH2 for the four gene pair orientations head–head, tail–tail, tail–head and head–tail based on all pairs of two directly adjacent genes in the data set by Lee *et al.* (2002).

transcription factor	head–head	tail–tail	tail–head	head–tail
ACE2	0.76	0.37	0.13	0.26
SWI5	0.80	0.26	0.12	0.20
FKH2	0.89	0.29	0.27	0.22

The correlations of ChIP-chip measurements of gene pairs in head–head orientation are significantly higher than in the three other categories.

context of the prediction of histone modifications from tiling array ChIP-chip data. Also ChIPmix (Martin-Magniette *et al.*, 2008) based on a linear regression mixture model can be applied to this kind of analysis.

Here, we study three methods for the prediction of transcription factor target genes from promoter array ChIP-chip data. We consider (i) a standard log-fold change (LFC) analysis that does not integrate dependencies between adjacent genes on DNA; (ii) a two-state HMM that models dependencies between adjacent genes on DNA; and (iii) an extension of the two-state HMM to an HMM with scaled transition matrices (SHMM) that specifically models directly adjacent genes on DNA that are in head–head orientation to each other. The three methods are applied to two datasets, one of the yeast *Saccharomyces cerevisiae* and another one of the model plant *Arabidopsis thaliana*, to directly compare their predicted target genes. Regarding the HMM approach, the two-state architecture follows the proposal of Li *et al.* (2005). Our approach is extended in that way that all HMM parameters are directly learned from the ChIP-chip data using a Bayesian version of the Baum–Welch algorithm described in Seifert *et al.* (2009). The concept of SHMMs is based on the key assumption that promoters of directly adjacent genes in head–head orientation on DNA tend to have more similar ChIP-chip measurements then directly adjacent genes in tail–tail, tail–head or head–tail orientations. That gene pair orientation specific correlations of ChIP-chip measurements exist is clearly shown in Table 1 for the three transcription factors ACE2, SWI5 and FKH2 studied in *Saccharomyces cerevisiae*, and in Figure 1 for the seed-specific transcription factor ABI3 analyzed in *Arabidopsis thaliana*. The high correlations of ChIP-chip measurements of promoters belonging to adjacent genes in head–head orientation are expected due to the design of the promoter array that contains spotted promoter fragments of each gene. Thus, depending on the distance between two genes in head–head orientation and the length of the hybridized DNA segments, one expects these correlations. The SHMM approach makes use of this observation by modeling that genes in head–head orientation have a higher probability that either both are targets of the transcription factor or both are non-targets of this transcription factor with respect to all other gene pair orientations. In general, good introductions to HMMs are given by Rabiner (1989) or Durbin *et al.* (1998). Extensions of standard HMMs with one transition matrix to HMMs with more than one transition matrix are described in Knab *et al.* (2003). Some more details to SHMMs can be found in Seifert (2006), and a concept similar to SHMMs has been developed by Meyer and Durbin (2004) with an application to gene prediction.

In this article, we focus on the analysis of two promoter array ChIP-chip datasets. We start with an initial study in the context of the cell cycle of *S.cerevisiae*. The three methods LFC, HMM and SHMM are used to predict common target genes bound by the transcription factors ACE2 and SWI5, and ACE2 and FKH2. We evaluate the common target genes using the Saccharomyces Genome Database by Cherry *et al.* (1997). Regarding *A.thaliana*, ChIP-chip based on promoter arrays was established for the seed-specific transcription factor ABI3 (ARABIDO-SEED, 2008). ABI3 is one of the fundamental regulators of seed development involved in controlling chlorophyll degradation, storage product accumulation, and desiccation tolerance (Mönke *et al.*, 2004; Suzuki *et al.*, 2003; To *et al.*, 2006; Vicente-Carbajosa and Carbonero, 2005). In an in-depth study, we use the three methods LFC, HMM and SHMM to identify putative ABI3 target genes, and we evaluate these genes using (i) publicly available expression data from Genevestigator (Hruz *et al.*, 2008; Zimmermann *et al.*, 2004) and (ii) transient assays, as described in Reidt *et al.* (2000), have been performed in wet laboratory experiments to test whether a promoter of a putative target gene is regulated by ABI3 or not.

2 METHODS

2.1 Yeast dataset

Publicly available promoter array ChIP-chip data from Lee *et al.* (2002) are used to identify common target genes of the cell cycle specific transcription factors ACE2 and SWI5, and ACE2 and FKH2. We downloaded the gene specific file from *http://web.wi.mit.edu/young/regulator_network* including the measured ratio r_t of immunoprecipitated DNA to input DNA for each promoter mapped to its corresponding gene t. For the transcription factors ACE2, SWI5 and FKH2, we extracted the measured gene specific ratios r_t for all genes t and transformed them into log-ratios $o_t = \log_2(r_t)$ for each of these three transcription factors. In addition to this, we mapped these log-ratios to their corresponding positions in the yeast genome using the Saccharomyces Genome Database (Cherry *et al.*, 1997). This leads to one ChIP-chip profile $o = o_1, \ldots, o_T$ per chromosome for each of the three transcription factors. The genome of the yeast *S.cerevisiae* consists of sixteen chromosomes, and due to that we obtain 16 ChIP-chip profiles for each transcription factor.

2.2 Arabidopsis dataset

The ChIP-chip technique by Ren *et al.* (2000) and Iyer *et al.* (2001) was applied to *A.thaliana* wild-type seeds to determine target genes of the ABI3 transcription factor. Isolated DNA fragments bound by ABI3 were amplified, radio-labeled, and hybridized to a macroarray containing 11904 promoters of *A.thaliana*. The corresponding control sample was obtained from the input chromatin of the wild-type 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 (Bolstad *et al.*, 2003) separately for the ABI3 ChIP-chip experiments and the input chromatin control experiments. In a second step, we combine each normalized ABI3 ChIP-chip experiment with its corresponding input chromatin control experiment by calculating the log-ratio $o_t = \log_2(I_{\text{ABI3}}(t)/I_{\text{INPUT}}(t))$ of immunoprecipitated DNA to input DNA for all genes t that are represented in the ABI3 ChIP-chip experiment by their promoter fragments on the macroarray. Here, $I_{\text{ABI3}}(t)$ is the normalized signal intensity of the promoter fragment of gene t in the ABI3 ChIP-chip experiment, and in analogy, $I_{\text{INPUT}}(t)$ represents the normalized signal intensity of the promoter fragment of gene t in the input chromatin control experiment. We map all 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 $o=o_1,...,o_T$ per chromosome. We obtain 25 ChIP-chip profiles, one for each of the five chromosomes for each of the five replicates.

2.3 Standard LFC analysis for target gene detection

The log-ratio of immunoprecipitated DNA to input DNA that is measured for a promoter characterizes the potential of the corresponding gene to be a target gene of the analyzed transcription factor. Thus, we expect that putative target genes have log-ratios that are significantly greater than zero. For each experiment an initial list is created that contains all gene identifiers of the ChIP-chip profiles in decreasing order of their log-ratios. That is, genes with log-ratios significantly greater than zero are at the top of this list. Considering the replicates of an experiment, we use the resulting lists to determine the intersection of the top k candidate genes in each list. This allows to assess the degree of reproducibility between the replicates of an experiment. All genes in the intersection are interpreted as putative target genes of the analyzed transcription factor.

2.4 HMM for target gene detection

2.4.1 HMM description We use a two-state HMM $\lambda=(S,\pi,A,E)$ with Gaussian emission densities for the genome-wide detection of putative target genes of a transcription factor. The basis of this HMM is the set of states $S=\{-,+\}$. State '$-$' corresponds to a gene that is not a target of the analyzed transcription factor, and state '$+$' corresponds to a gene that is a target of this transcription factor. We denote the state of gene 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 gene t is assumed to be drawn from a Gaussian emission density with mean and standard deviation depending on state q_t. We denote the vector of emission parameters by $E=(\mu_-,\mu_+,\sigma_-,\sigma_+)$ with means μ_- and μ_+, and standard deviation 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 $i\in S$.

2.4.2 HMM initialization In general, an initial HMM should distinguish putative target genes of the analyzed transcription factor from non-target genes with respect to their log-ratios in the ChIP-chip profile. Hence, a histogram of log-ratios helps to find good initial HMM parameters. The choice of initial parameters addresses the presumptions that the proportion of non-target genes is much higher than that of target genes, and that the number of successive non-target genes is also much higher than the number of successive target genes. In our case studies, 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 specific 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.

2.4.3 HMM training We train the initial HMM λ^1 based on all ChIP-chip profiles using a maximum *a posteriori* (MAP) variant of the standard Baum–Welch algorithm (Baum, 1972; Durbin *et al.*, 1998; Rabiner, 1989). This algorithm is part of the class of Expectation Maximization (EM) algorithms (Dempster *et al.*, 1977) 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-Inverted-Gamma priors for emission parameters E with hyper-parameters $\eta_-=0$ and $\eta_+=2$ (a priori means), $\epsilon_-=\epsilon_+=10^3$ (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. In the case of data with mean log-ratio of about zero, the choice of these prior parameters can be simplified by using the a priori mean $\eta_-=0$ for the non-target gene state in combination with a user-defined a priori mean $\eta_+\in\mathbb{R}^+$ for the target gene state. However, to provide the full flexibility the user can also make own settings of the prior parameters with respect to their influences on the HMM parameters during the MAP training, which can be derived from Seifert *et al.* (2009). 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 $<10^{-9}$. More details to the MAP training and the incorporated prior are given in Seifert *et al.* (2009).

2.4.4 HMM target gene detection The state '$+$' of the trained HMM λ models the potential of genes to be targets of the analyzed transcription factor. To quantify this potential, we calculate the probability $\gamma_t(+)=P[Q_t=+|O=o,\lambda]$ of being a target gene for each gene t within a ChIP-chip profile o. This state posterior of state '$+$' is computed using the Forward–Backward procedures of HMMs (Durbin *et al.*, 1998; Rabiner, 1989). For each experiment, we create a list that contains all gene identifiers of the ChIP-chip profiles in decreasing order of their state posteriors $\gamma_t(+)$. Considering the replicates of an experiment, we use these lists to determine the intersection of the top k candidate genes of each list. In analogy to the standard LFC approach, we interpret all genes in the intersection as putative target genes of the analyzed transcription factor.

2.5 SHMM for target gene detection

2.5.1 SHMM description The general concept of SHMMs allows to analyze ChIP-chip profiles in the context of orientations of adjacent genes on the DNA. Two directly adjacent genes 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 special importance for the analysis of promoter array ChIP-chip data. In this orientation, the two corresponding genes have the potential to share a common promoter region depending on the distance between these genes (Fig. 1). This fact in combination with the observation that the log-ratios of promoters of genes in head–head orientation show significantly higher correlations, as shown in Figure 1 and in Table 1, compared with 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 exhibit the same status based on the log-ratios measured for their promoter fragments. In comparison to tail–tail, tail–head and head–tail orientations, it is more likely that both genes of a head–head orientation are either target genes or non-target genes of the analyzed transcription factor. For this reason, we assign to each pair of successive genes t and $t+1$ on a chromosome one gene pair orientation class $c(d_t)$ depending on the orientation of both genes to each other in combination with the chromosomal distance d_t of theses two genes. The gene pair orientation class of successive genes 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 this information into a two-state SHMM $\lambda=(S,\pi,A,\vec{f},E)$ with two gene pair orientation classes to detect putative target genes of the analyzed transcription factor. The parameters S, π and E are defined as described in the previous section of the HMM approach. 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 $A=(A_1,A_2)$. These two transition matrices distinguish head–head gene pairs from others by scaling the basic transition probabilities $a_{ii}\in(0,1)$ and $a_{ij}=1-a_{ii}$ for $i,j\in S$ with $i\neq j$ using the vector of scaling factors $\vec{f}=(f_1:=1,f_2)$ with $f_2\in\mathbb{R}^+$ and $f_2>f_1$. This results in transition matrix

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

for gene pair orientation class $l\in\{1,2\}$. The expected state duration of state $i\in S$ in A_1 is scaled from $1/(1-a_{ii})$ to $f_2/(1-a_{ii})$ in A_2. 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 gene 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 gene pair modeled by A_2 it is more likely that both genes are targets or non-targets of the analyzed transcription factor compared to other gene pairs modeled by A_1. The log-ratios of genes are modeled as described in the HMM approach.

2.5.2 SHMM initialization The basic initialization of the SHMM is identical to that of the HMM. In addition to that, we must choose a distance threshold b for the gene pair orientation classes and a scaling factor f_2 to specify the degree of differentiation between head–head orientations modeled by A_2 and all other orientations modeled by A_1. For the initial study in yeast, we neglect b by setting it to ∞, because most of the genes have distances less than 2 Kb to its next adjacent gene on DNA, and because the correlations of ChIP-chip measurements of head–head gene pairs shown in Table 1 are generally high for all transcription factors. Additionally, we consider the scaling factor $f_2=4.0$. In the in-depth ABI3 case study, we always use $b=9$ Kb motivated by Figure 1, because at greater chromosomal distance the correlations of ChIP-chip measurements of head–head gene pairs do not significantly differ from other gene pairs. In addition to this, we assess all values of f_2 in the interval 1.1 to 10 in steps of 0.1.

2.5.3 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 for the estimation of their transition matrices. Details of the parameter estimation are described in Seifert (2006).

2.5.4 SHMM target gene detection The putative target genes of the analyzed transcription factor are determined in analogy to the HMM approach. The calculation of the state posterior $\gamma_t(+)$ is now done with respect to the gene pair orientation classes using the Forward–Backward procedures of HMMs.

3 RESULTS AND DISCUSSION

In this section, we first make an initial study to compare the three approaches LFC, HMM and SHMM based on their prediction of common target genes of the yeast cell cycle transcription factors ACE2 and SWI5, and ACE2 and FKH2. Subsequent to this, we focus on an in-depth study of LFC, HMM and SHMM for predicting target genes of the seed-specific transcription factor ABI3 in *A.thaliana*.

3.1 Yeast dataset

3.1.1 Common target genes of yeast cell cycle regulators The publicly available promoter array ChIP-chip dataset by Lee *et al.* (2002) provides the opportunity to predict common target genes of the yeast cell cycle transcription factors ACE2 and SWI5, and ACE2 and FKH2 by the three methods LFC, HMM and SHMM. The transcription factors ACE2 and SWI5 are known to regulate common genes expressed at the boundary of the M/G1 phase of the cell cycle (Lee *et al.*, 2002; Mc Bride *et al.*, 1999), and the transcription factors ACE2 and FKH2 control the regulation of a common set of genes in the G1 phase of the cell cycle (Lee *et al.*, 2002). We use the LFC method to score the putative target genes of ACE2, SWI5 and FKH2 separately based on the log-ratios of the genes in the corresponding ChIP-chip profiles of the three transcription factors. Subsequent to this, we determine the intersection of the top 75 scoring genes for ACE2 and SWI5, and of the top 130 scoring genes for ACE2 and FKH2. This ensures that all putative common target genes have mean log-ratios greater than one. To motivate the application of the HMM and the SHMM approach, the Pearson correlations of the ChIP-chip measurements for the four gene pair orientations head–head, tail–tail, tail–head and head–tail are shown in Table 1 for the three transcription factors ACE2, SWI5 and FKH2. We observe positive correlations in all four categories, which motivates the usage of the HMM for predicting common target genes. Additionally, the correlations in the head–head category are significantly higher in comparison to the three other categories. This observation provides the basics for using the SHMM to specifically model the head–head orientations for the prediction of common transcription factor target genes. For the comparison to the LFC method, we separately train an HMM and a SHMM with scaling factor $f_2=4.0$ on all ChIP-chip profiles of ACE2 and SWI5. Separately for the HMM and the SHMM, we determine putative common target genes by computing the intersection of the top 75 scoring genes. In analogy to this, we determine putative common target genes of ACE2 and FKH2 at the level of the top 130 target genes. The results for the comparison of LFC, HMM and SHMM are shown in Figure 2. In both cases, all common target genes predicted by LFC and HMM have also been predicted by the SHMM. Additionally, the SHMM predicted two putative target genes that have not been identified by LFC and HMM. To investigate whether the putative common target genes that have only been predicted by the SHMM, or together by the HMM and the SHMM are involved in the regulation of the yeast cell cycle we used the Saccharomyces Genome Database (Cherry *et al.*, 1997). Regarding the common target genes of ACE2 and SWI5, the gene YJL160C has only been predicted by the SHMM. This gene is a member of the PIR family of cell wall proteins with functions in sporulation, and its gene expression level is weakly cell cycle regulated peaking in the M phase of the cell cycle (de Lichtenberg *et al.*, 2005; Enyenihi and Saunders, 2003; Giaver *et al.*, 2002; Jung and Levin, 1999). Currently, no function is known for gene YBR157C, which has been predicted by the SHMM and the HMM. Considering the common target genes of ACE2 and FKH2, the gene YER127W has only been predicted by the SHMM. This gene encodes a protein which is essential for the maturation of the 18S rRNA. Repression of the gene expression of this gene leads to an abnormal progression of the G1 phase of the cell cycle (Yu *et al.*, 2006). The genes YER126C, YFL021W and YFL022C have been identified as putative common target genes of ACE2 and FKH2 by the HMM and by the SHMM. The protein of gene YER126C is part of the 66S pre-ribosomal particles and contributes to the processing of the 27S pre-rRNA. The overexpression of this gene leads to a decrease in the vegetative growth of the yeast (Horsey *et al.*, 2004), which has consequences for the G1 phase of the cell cycle where the cell grows. The gene YFL021W encodes a transcription factor that activates genes involved in nitrogen catabolite repression. The gene YFL022C encodes the alpha subunit of the cytoplasmic

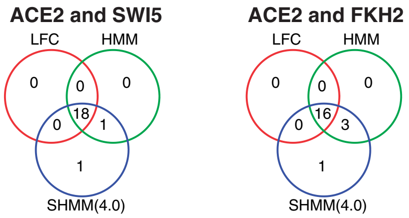

Fig. 2. Venn diagrams for comparing the prediction of common target genes of the *S.cerevisiae* cell cycle transcription factors ACE2 and SWI5, and ACE2 and FKH2 by the three methods LFC, HMM and the SHMM(4.0) that specifically models the head–head orientation of adjacent gene pairs. In both cases, the SHMM(4.0) is the most general model that predicted the highest number of putative common target genes including all target genes predicted by LFC and HMM.

phenylalanyl-tRNA synthetase. The overexpression of this gene is known to lead to a delay or an arrest of the G2 or M phase of the cell cycle (Niu *et al.*, 2008). In summary, for both pairs of transcription factors ACE2 and SWI5, and ACE2 and FKH2 the SHMM approach predicted the highest number of putative common target genes that are involved in the regulation of the yeast cell cycle.

3.2 Arabidopsis dataset

3.2.1 Differences between HMM and SHMMs for ABI3 The HMM approach allows to analyze ChIP-chip data in the context of chromosomal locations of genes. The application of SHMMs extends this analysis by discriminating different types of gene pair orientations. In this study, we investigate how SHMMs behave in comparison to the standard HMM. For that reason, the Viterbi algorithm (Durbin *et al.*, 1998; Rabiner, 1989; Viterbi, 1967) is used 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 Viterbi annotation of a gene t with log-ratio o_t is given by $q_t \in S$, which we interpret as the promoter of gene t is either a putative ABI3 target or not. The scaling factor allows to directly influence the annotation behavior for head–head gene pairs. That is, the higher f_2 the more likely it is that both genes of such head–head pairs are either putative ABI3 targets or not, and the closer f_2 is set to one the more similar the annotation behavior of the SHMM gets to that of the HMM. The results are summarized in Figure 4. As expected, we observe that the number of head–head gene pairs for which both genes of such a pair have identical annotations increases for increasing scaling factor f_2, and consequently, the number of head–head gene pairs for which both genes of such a pair have different annotations decreases. Obviously, each change in the annotation of a head–head gene pair leads either to a change in the annotation of the upstream, downstream, or both of these gene pairs. We find that the number of non-head–head gene pairs for which both genes of such a pair are annotated as putative ABI3 targets decreases only slightly for SHMMs with increasing scaling factor f_2 compared with the HMM. We clearly observe substantially more decrease in the number of non-head–head gene pairs for which both genes of such a pair are annotated as putative non-target genes for SHMMs with increasing scaling factor f_2 in relation to the HMM. Consequently, the number of non-head–head gene pairs for which both genes of such a pair have different annotations increases with increasing scaling factor f_2. This comparative study points out that the Viterbi 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.

3.2.2 Comparison of LFC, HMM and SHMM for the prediction of ABI3 target genes We use the LFC method for scoring putative ABI3 target genes based on the log-ratios of the genes. This method neglects chromosomal locations and gene pair orientations. For comparison, we make use of the HMM that models chromosomal locations of genes, and we make use of the SHMM that extends the HMM by modeling orientations of gene pairs. Thereby, both HMM approaches score putative ABI3 target genes via the state posterior of the target gene state. In this comparative 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 top list 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 quite different from the standard HMM (Fig. 4), and because the comparison of this model to SHMMs with higher scaling factors $f_2 = 6$ and $f_2 = 10$ yielded identical target genes. For each approach, we score all five ChIP-chip experiments to determine the intersection of putative ABI3 target genes for the top 50, 100, 150 and 200 candidates obtained from these experiments. Then, we use Venn diagrams to directly compare the candidate genes for these four top lists given by all three methods. The results are shown in Figure 3a. We observe that the SHMM predicted the highest number of putative ABI3 target genes, whereas the LFC method predicted the smallest number. Comparing the Venn diagrams of 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 in the transition of the top 150 list to the top 200 list are completely predicted by the SHMM. Next, we investigate whether the putative ABI3 target genes that have only been predicted by the SHMM at the level of the top 200 candidates are the consequence of specifically modeling the head–head orientations. For that purpose, we also trained a SHMM that specifically models tail–tail orientations using the identical initial settings. Figure 3b shows that the SHMM that specifically models tail–tail orientations has a prediction behavior that is nearly identical to that of the standard HMM with perfect agreement at the level of the top 50 and 150 candidates, and one additional putative target gene at the level of the top 200 candidates. This coincides with the observation shown in Figure 1 that the measured log-ratios of gene pairs in tail–tail orientation tend to be uncorrelated. Due to that, the specific modeling of tail–tail orientations has nearly no effect on the prediction of putative ABI3 target genes. Figure 3c shows that the prediction results of the SHMM that specifically models tail–tail orientations are completely included in the set of predicted putative ABI3 target genes of the SHMM that specifically models head–head orientations. This indicates that the gain of additional putative ABI3 target genes is based on the specific modeling of head–head orientations. In summary, this emphasizes that the SHMM approach that models head–head orientations tends to be more general in the prediction of putative ABI3 target genes than the HMM, the LFC and the SHMM that models tail–tail orientations.

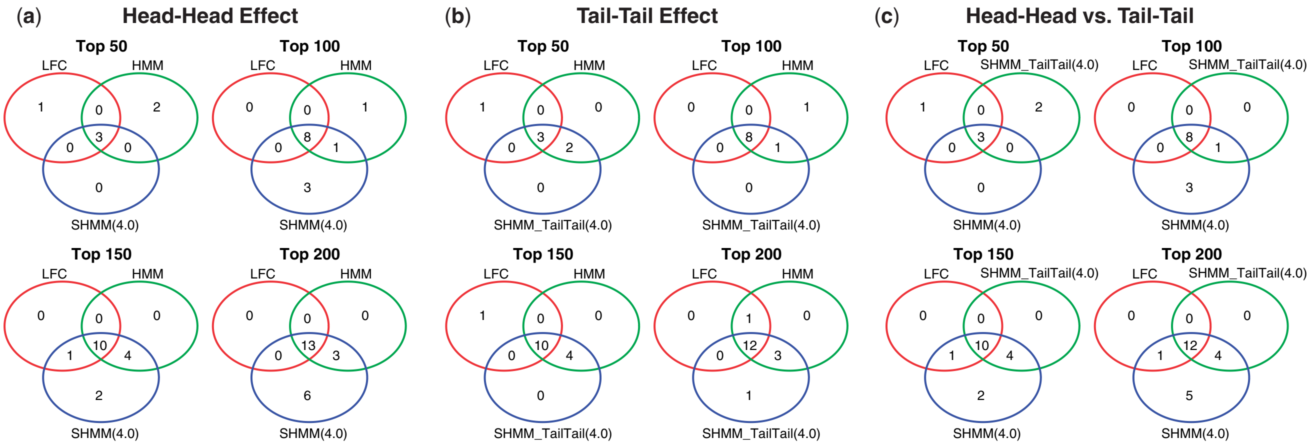

Fig. 3. Venn diagrams for comparing the number of predicted putative ABI3 target genes by the standard LFC, the standard HMM, the SHMM(4.0) that specifically models head–head orientations of genes, and the SHMM_TailTail(4.0) that specifically models the tail–tail orientation of genes for validating the SHMM(4.0). **(a)** Venn diagrams that show the comparison of the number of predicted putative ABI3 target genes by LFC, HMM and the SHMM(4.0) that specifically models the head–head orientations. The SHMM(4.0) is the most general model that predicted the highest number of putative target genes including all genes found by LFC and HMM at the level of the top 200 candidates. **(b)** Venn diagrams that show the comparison of predicted putative ABI3 target genes by LFC, HMM and the SHMM_TailTail(4.0) that specifically models tail–tail orientations. The SHMM_TailTail(4.0) does predictions nearly identical to the HMM with perfect agreement at the level of the top 50 and 150 candidates. The total number of predicted putative ABI3 target genes is less than in Figure 3a. **(c)** Venn diagrams that show the comparison of predicted putative ABI3 target genes by LFC, SHMM_TailTail(4.0) and the SHMM(4.0). The SHMM(4.0) that specifically models the head–head orientation is the most general model that predicted the highest number of putative ABI3 target genes including all genes predicted by LFC and the SHMM_TailTail(4.0) that specifically models tail–tail orientations. This states that the gain of additional putative ABI3 target genes is based on the specific modeling of head–head orientations by the SHMM(4.0).

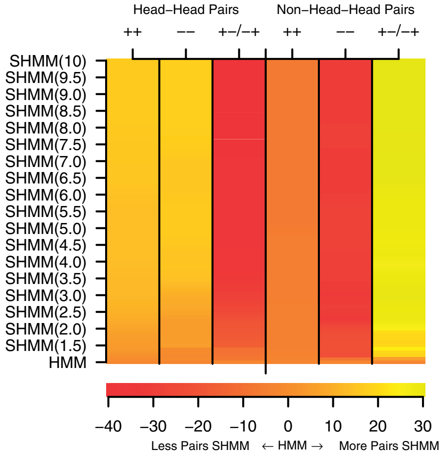

Fig. 4. Differences of gene pair annotations of the trained SHMM(f_2) with scaling factor $f_2 \in [1.1, 10]$ in steps of 0.1 in relation to the trained HMM based on Viterbi annotations. The HMM is encoded by the orange shade with value zero. The annotations $++$, $--$ and $+-/-+$ of gene pairs define that either both genes are putative targets, non-targets or only one gene of both is a putative target of ABI3.

3.2.3 Gene expression analysis of putative ABI3 target genes Next we investigate how putative target genes are regulated by ABI3. For that purpose, we use Genevestigator (Zimmermann *et al.*, 2004) as an 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 in wet laboratory experiments to test whether the promoters of putative ABI3 target genes in fusion with the glucuronidase (GUS) reporter gene react on ABI3. The results are shown in Table 2. 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 1000 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. The promoter of gene T21 shows nearly a 2-fold repression of the GUS expression, which is not reflected by its Genevestigator score. Interestingly, the genes T21 and T22 are in head–head orientation to each other, and thus 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 three of these 15 target genes activated by ABI3 and the one target gene repressed by ABI3. In contrast to these 16 target genes, the six 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 1000 randomly selected genes.

Table 2. Overview of predicted ABI3 target genes by LFC, HMM and SHMM(4.0) at the level of the top 200 candidates in Figure 3a.

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

The ID column contains anonymized target gene identifiers (a manuscript discussing these genes is currently in preparation). The numbers 1 and 0 in the method columns LFC, HMM and SHMM(4.0) encode whether a gene is predicted (1) or missed (0). GV (Genevestigator) quantifies the gene expression of a target gene within the categories inflorescence, silique and seed as described in Section 3.2.3. TA (Transient assay) contains the measured fold-change of the GUS gene expression for a target gene promoter under ABI3 expression in relation to this target gene promoter lacking the expression of ABI3.

Interestingly, five of these six putative target genes are in head–head orientation to one of the previous target genes activated by ABI3. Next, we address the question if these six putative ABI3 target genes are also under control of ABI3. To test this hypothesis, the promoters of four of these six 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 2-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 promoter fragment on the promoter arrays. The Genevestigator score of T23 is significantly higher than those of the 1000 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 these results if the underlying test system is capable of simulating the natural situation in seeds. In total, 20% of the target genes with high Genevestigator scores and activation by ABI3 could be predicted only through the application of the SHMM approach and would have been missed using the LFC or HMM approach. This points out the relevance of SHMMs for the detection of ABI3 target genes.

4 CONCLUSIONS

We studied the LFC, HMM and new SHMM approach for the analysis of promoter array ChIP-chip data of the yeast *S.cerevisiae* and of the model plant *A.thaliana*. The application of HMMs and SHMMs to the analysis of promoter array ChIP-chip data is motivated by the observation of positive correlations of ChIP-chip measurements of directly adjacent genes on DNA as shown in Table 1 for *S.cerevisiae*, or in Figure 1 for *A.thaliana*. Especially, the new SHMM approach takes additionally into account that the ChIP-chip measurements of the head–head gene pairs have significantly higher correlations than those of all other gene pairs. Regarding all three methods, the SHMM predicted the highest number of target genes for the promoter array ChIP-chip data sets of *S.cerevisiae* and *A.thaliana* including all target genes of LFC and HMM. However, the number of predicted target genes is not an optimal criterion to decide which of the methods should be preferred. For that reason, we searched in literature and data bases, analyzed expression data, and had access to the results of transient assays to validate the prediction results. Using these independent sources for validation, the SHMM showed the best performance of all three methods for the promoter array ChIP-chip datasets of *S.cerevisiae* and *A.thaliana*. Taking this together, this indicates that the SHMM approach is a valuable tool that could also be applied to other promoter array ChIP-chip data sets.

ACKNOWLEDGMENTS

We thank the groups of Lothar Altschmied, Helmut Bäumlein and Udo Conrad at the IPK, and especially Urs Hähnel and Gudrun Mönke for the ABI3 ChIP–chip data, transient assays and valuable discussions. Regarding the development of the basic concept underlying the SHMM approach, we thank Alexander Schliep and Stefan Posch. We thank the anonymous reviewers for their valuable comments. We thank the organizers of the GCB 2008 in Dresden for the possibility to present this work.

Funding: BMBF (grants 0312706A and 0313155); Ministry of culture Saxony-Anhalt (grant XP3624HP/0606T).

Conflict of Interest: none declared.

REFERENCES

ARABIDO-SEED. (2008) A trilateral project between France, Spain, and Germany studying seed development of Arabidopsis thaliana. available at http://arabidoseed.ipk-gatersleben.de (last accessed date 2009).

Baum,L.E. (1972) An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. *Inequalities*, **3**, 1–8.

Bolstad,B.M. *et al.* (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.

Cherry,J.M. *et al.* (1997) Genetic and physical maps of Saccharomyces cerevisiae. *Nature*, **387(Suppl 6632)**, 67–73.

Chung,H.-R. *et al.* (2007) A physical model for tiling array analysis. *Bioinformatics*, **23 ISMB/ECCB 2007**, i80–i86.

de Lichtenberg,U. *et al.* (2005) New weakly expressed cell cycle-regulated genes in yeast. *Yeast*, **22**, 1191–1201.

Dempster,A. *et al.* (1977) Maximum likelihood from incomplete data via the EM algorithm. *J.R Stati. Soc., Ser. B*, **39**, 1–38.

Durbin,R. *et al.* (1998) *Biological Sequence Analysis - Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge.

Enyenihi,A.H. and Saunders,W.S. (2003) Large-scale functional genomic analysis of sporulation and meiosis in Saccharomyces cerevisiae. *Genetics*, **163**, 47–54.

Giaver,G. *et al.* (2002) Functional profiling of the Saccharomyces cerevisiae genome. *Nature*, **418**, 387–391.

Horsey,E.L. *et al.* (2004) Role of the yeast Rrp1 protein in the dynamics of pre-ribosome maturation. *RNA*, **10**, 813–827.

Hruz,T. *et al.* (2008) Genevestigator V3: a reference expression database for the meta-analysis of transcriptomes. *Adv. Bioinform*, **2008**, doi:10.1155/2008/420747.

Humburg,P. *et al.* (2008) Parameter estimation for robust HMM analysis of ChIP-chip data. *BMC Bioinformatics*, **9**, doi:10.1186/1471-2105-9-343.

Iyer,V.R. *et al.* (2001) Genomic binding sites of the yeast cell-cycle transcription factors SFB and MBF. *Nature*, **409**, 533–538.

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

Johnson,D.S. *et al.* (2008) Systematic evaluation of variability in ChIP-chip experiments using predefined DNA targets. *Genome Res*, **18**, 393–403.

Jung,U.S. and Levin,D.E. (1999) Genome-wide analysis of gene expression regulated by the yeast cell wall integrity signalling pathway. *Mol Microbiol*, **34**, 1049–1057.

Keles,S. *et al.* (2004) Multiple testing methods for ChIP-chip high density oligonucleotide array data. *Working Paper Series 147*. U.C. Berkeley Division of Biostatistics, University of California, Berkeley, CA.

Knab,B. *et al.* (2003) Model-based clustering with Hidden Markov Models and its application to financial time-series data. In M. *Schader,M. et al. (eds), Between Data Science and Applied Data Analysis. Springer*, pp. 561–569.

Lee,T.I. *et al.* (2002) Transcripitonal Regulatory Networks in Saccaromyces cerevisiae. *Science*, **298**, 799–804.

Li,W. *et al.* (2005) A hidden Markov model for analyzing ChIP-chip experiments on genome tiling arrays and its application to p53 binding sequences. *Bioinformatics*, **21**, i274–i282.

Martin-Magniette,M.-L. *et al.* (2008) ChIPmix: Mixture model of regressions for two-color ChIP-chip analysis. *Bioinformatics*, **24 ECCB 2008**, i181–i186.

Mc Bride,H.J. *et al.* (1999) Distinct regions of the Swi5 and Ace2 transcription factors are required for specific gene activation. *J. Biol. Chem*., **274**, 21029–21036.

Meyer,I.M. and Durbin,R. (2004) Gene structure conservation aids similarity based gene prediction. *Nucleic Acids Res.*, **32**, 776–783.

Mönke,G., *et al.* (2004) Seed-specific transcription factors ABI3 and FUS3: molecular interaction with DNA. *Planta*, **219**, 158–166.

Niu,W. *et al.* (2008) Mechanisms of cell cycle control revealed by a systematic and quantitative overexpression screen in S. cerevisiae. *PLoS Genet*, **4**, e1000120.

Rabiner,L. (1989) A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proc. IEEE*, **77**, 257–286.

Reidt,W. *et al.* (2000) Gene regulation during late embryogenesis: the RY motif of maturation-specific gene promoters is a direct target of the FUS3 gene product. *Plant J.*, **21**, 401–408.

Ren,B. *et al.* (2000) Genome-Wide Location and Function of DNA Binding Proteins. *Science*, **290**, 2306–2309.

Seifert,M. (2006) Analysing microarray data using homogeneous and inhomogeneous hidden Markov models. Diploma Thesis. Martin Luther University.

Seifert,M. *et al.* (2009) Array-based genome comparison of Arabidopsis ecotypes using hidden Markov models. In *Proceedings of the Biosignals 2009*, Portugal, pp. 3–11.

Suzuki,M. *et al.* (2003) Viviparous Alters Global Gene Expression Patterns through Regulation of Abscisic Acid Signaling. *Plant Physiol.*, **132**, 1664–1677.

To,A. *et al.* (2006) A Network of Local and Redundant Gene Regulation Governs Arabidopsis Seed Maturation. *Plant Cell*, **18**, 1642–1651.

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

Viterbi,A.J. (1967) Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans. Inform. Theory*, **13**, 260–269.

Yu,L. *et al.* (2006) A survey of essential gene function in the yeast cell division. *Mol. Biol. Cell*, **17**, 4736–4747.

Zimmermann,P. *et al.* (2004) GENEVESTIGATOR. Arabidopsis microarray database and analysis toolbox. *Plant Physiol.*, **136**, 2621–2632.