# ARRAY-BASED GENOME COMPARISON OF ARABIDOPSIS ECOTYPES USING HIDDEN MARKOV MODELS

Michael Seifert[1], Ali Banaei[1], Jens Keilwagen[1], Michael Florian Mette[1], Andreas Houben[1]
François Roudier[2], Vincent Colot[2], Ivo Grosse[3] and Marc Strickert[1]
[1]*Leibniz Institute of Plant Genetics and Crop Plant Research, Corrensstr. 3, 06466 Gatersleben, Germany*
[2]*Ecole Normale Supérieure, Département de Biologie, CNRS UMR8186, 46 rue d'Ulm, 75230 Paris cedex 05, France*
[3]*Martin Luther University, Institute of Computer Science, Von-Seckendorff-Platz 1, 06120 Halle, Germany*

Keywords: Array-CGH, Comparative Genomics, *Arabidopsis* Ecotypes, Hidden Markov Model(HMM).

Abstract: *Arabidopsis thaliana* is an important model organism in plant biology with a broad geographic distribution including ecotypes from Africa, America, Asia, and Europe. The natural variation of different ecotypes is expected to be reflected to a substantial degree in their genome sequences. Array comparative genomic hybridization (Array-CGH) can be used to quantify the natural variation of different ecotypes at the DNA level. Besides, such Array-CGH data provides the basics to establish a genome-wide map of DNA copy number variation for different ecotypes. Here, we present a new approach based on Hidden Markov Models (HMMs) to predict copy number variations in Array-CGH experiments. Using this approach, an improved genome-wide characterization of DNA segments with decreased or increased copy numbers is obtained in comparison to the routinely used segMNT algorithm. The software and the data set used in this case study can be downloaded from http://dig.ipk-gatersleben.de/HMMs/ACGH/ACGH.html.

## 1 INTRODUCTION

The method of array-based comparative genomic hybridization (Array-CGH) has been widely applied to several genomes for studying deletions, insertions, and amplifications of DNA segments (Mantripragada et al., 2004) including studies on *Arabidopsis thaliana* (Borevitz et al., 2003; Martienessen et al., 2005; Fan et al., 2007) an important model organism in plant biology. Due to the broad geographic distribution of *Arabidopsis thaliana* ecotypes their the natural variation is expected to be reflected to a substantial degree in their genome sequences. The application of Array-CGH to these genomes allows to quantify the natural variation at the DNA level. The obtained Array-CGH data provides basics to establish a genome-wide map of DNA copy number variations between different ecotypes. Based on such a map, future studies of DNA-histone interactions, histone modifications, or transcript profiling will allow an improved comparison of different ecotypes.

One important bioinformatics tasks is to create a genome-wide map characterizing regions of DNA copy number variations in Array-CGH data of different ecotypes. In recent years, the prediction of DNA copy number variations in tumor data has received most attention, leading to the development of many different approaches for determining copy number variations in Array-CGH data. These approaches include genetic local search algorithms (Jong et al., 2004), adaptive weights smoothing, (Hupé et al., 2004), and Hidden Markov Models (HMMs) (Fridlyand et al., 2004; Marioni et al., 2006; Cahan et al., 2008). Contributions to the comparison of different approaches have been made by two recent studies (Lai et al., 2005; Willenbrock and Fridlyand, 2005).

The basic concept of applying HMMs to the analysis of Array-CGH data was initially developed by (Fridlyand et al., 2004). In this paper, we propose a new method based on HMMs for the detection of DNA segments with decreased or increased copy numbers from Array-CGH data. This approach has the following features: (i) we use a three-state HMM partitioning DNA segments into segments of decreased, unchanged, or increased copy numbers, (ii) we incorporate *a priori* knowledge into the training of the HMM, and (iii) we use permuted Array-CGH data to score predicted DNA segments with decreased or increased copy numbers. We apply this HMM ap-



Seifert M., Banaei A., Keilwagen J., Mette M., Houben A., Roudier F., Colot V., Grosse I. and Strickert M. (2009).
ARRAY-BASED GENOME COMPARISON OF ARABIDOPSIS ECOTYPES USING HIDDEN MARKOV MODELS.
In *Proceedings of the International Conference on Bio-inspired Systems and Signal Processing*, pages 3-11
DOI: 10.5220/0001123700030011

proach to Array-CGH data of *Arabidopsis thaliana* ecotypes from whole-genome NimbleGen tiling arrays. We obtain an improved genome-wide map of copy number variations compared to the standard segMNT algorithm (Roche NimbleGen, Inc., 2008) routinely used for this task.

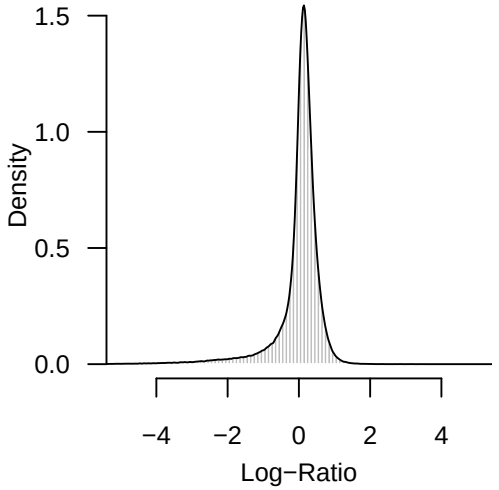


Figure 1: Histogram of log-ratios for *Arabidopsis thaliana* ecotype C24 versus ecotype Columbia. The majority of log-ratios is located at about zero, indicating that the majority of DNA segments have not changed their copy numbers in C24 compared to Columbia. Additionally, the proportion of log-ratios in the negative tail is significantly higher than the proportion of log-ratios in the positive tail. That is, we expect to find more DNA segments with decreased copy numbers in C24 than segments with increased copy numbers. The main reason for the asymmetry of the log-ratio distribution is that DNA segments which are deleted in the genome of Columbia, but which are present in the genome of C24 cannot be measured using a tiling array representing the reference genome of Columbia.

# 2 METHODS

## 2.1 Array-CGH Data

For the Array-CGH-based genome comparison of *Arabidopsis thaliana* ecotypes C24 and Columbia, leaf tissue is used to extract the DNA. Then the DNA is sheared by sonication and resulting DNA segments are differentially color-labeled for each ecotype. Subsequently, these DNA segments are hybridized to NimbleGen tiling arrays representing the reference genome of ecotype Columbia. The arrays are read out and further processed using the NimbleScan software (Roche NimbleGen, Inc.) resulting in normalized log-ratios $o_t = \log_2(I_t(\text{C24})/I_t(\text{Columbia}))$ for all tiles $t$ of an array, where $I_t(\text{C24})$ and $I_t(\text{Columbia})$ are the intensities of tile $t$ under the corresponding ecotype. Based on information about the chromosomal locations of all tiles of an array, we create an Array-CGH profile $o = o_1,\ldots,o_T$ for each chromosome where all log-ratios $o_t$ are represented in increasing order of their chromosomal positions. Each Array-CGH experiment consists of two independent arrays with tiles at slightly different chromosomal locations. That is, for two adjacent tiles of a chromosome spotted on one array there generally exists one tile on the other array having its chromosomal position between these first two tiles (interleaved design). The mean distance of two adjacent tiles on a chromosome is about 350 bp for one array. The length of hybridized segments is about 300 to 900 bp. Each array contains about 365,000 tiles that we separate into $K = 7$ Array-CGH profiles $o^1,\ldots,o^K$. The first five profiles represent the chromosomes of Arabidopsis thaliana and the last two contain the measurements for chloroplastic and mitochondrial DNA. We treat both arrays of an Array-CGH experiment independently to validate the reproducibility of our results. A histogram of log-ratios of the Array-CGH data used in this case study is shown in Fig. 1.

## 2.2 HMM-based Data Analysis

### 2.2.1 HMM Description

We use a three-state HMM $\lambda = (S,\pi,A,E)$ with Gaussian emission densities for the genome-wide detection of regional DNA copy number variations in *Arabidopsis thaliana* ecotypes. The basis of this HMM is the set of states $S = \{-,=,+\}$. These states model the copy number status of DNA regions in ecotype C24 that is compared to the reference genome of ecotype Columbia. Thus, state $-$ corresponds to DNA regions with decreased copy number, state $=$ corresponds to DNA regions with unchanged copy number, and state $+$ corresponds to DNA regions with increased copy number. The state of tile $t$ is denoted by $q_t \in S$. We assume that a state sequence $q = q_1,...,q_T$ belonging to an Array-CGH profile $o$ is generated by a homogeneous Markov model of order one with (i) start distribution $\pi = (\pi_-,\pi_=,\pi_+)$, where $\pi_i$ denotes the probability that the first state $q_1$ is equal to $i \in S$, and (ii) stochastic transition matrix

$$A = \begin{pmatrix} a_{--} & a_{-=} & a_{-+} \\ a_{=-} & a_{==} & a_{=+} \\ a_{+-} & a_{+=} & a_{++} \end{pmatrix}$$

where $a_{ij}$ denotes the conditional probability that state $q_{t+1}$ is equal to $j \in S$ given that state $q_t$ is equal to $i \in S$. Clearly, the start distribution fulfills $\sum_{i\in S}\pi_i = 1$,

and the transition probabilities of each state $i \in S$ fulfill $\sum_{j \in S} a_{ij} = 1$. The state sequence is assumed to be non-observable, i.e. hidden, and the log-ratio $o_t$ of tile $t$ 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 $E = (\mu_-, \mu_=, \mu_+, \sigma_-, \sigma_= \sigma_+)$ with mean $\mu_i \in \mathbb{R}$ and standard deviation $\sigma_i \in \mathbb{R}^+$ for the Gaussian emission density

$$b_i(o_t) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{1}{2}\frac{(o_t - \mu_i)^2}{\sigma_i^2}\right)$$

of log-ratio $o_t$ given state $q_t = i \in S$. An illustration of the proposed HMM is given in Fig. 2. Since we have nearly equidistant tiles along a chromosome with distances about 350 bp for adjacent tiles, we do not model chromosomal distances between adjacent tiles. Approaches that explicitly take adjacent distances into account are e.g. BioHMM (Marioni et al., 2006) and RJaCGH (Rueda and Díaz-Uriate, 2007).

### 2.2.2 HMM Initialization

In general, a good initial HMM should differentiate DNA regions of decreased or increased copy numbers from DNA regions of unchanged copy numbers with respect to their log-ratios in the Array-CGH profile. Hence, a histogram of log-ratios, like in Fig. 1, helps to find good initial HMM parameters. The choice of initial parameters addresses the two realistic presumptions. The first one is that the proportion of DNA regions with unchanged copy numbers is much higher than that of DNA regions of decreased or increased copy numbers. The second presumption is that the number of successive tiles with unchanged DNA copy numbers is also much higher than the number of successive tiles with decreased or increased DNA copy numbers. In this case study, we use $\pi_- = 0.2$, $\pi_= = 0.75$, and $\pi_+ = 0.05$ resulting in an initial start distribution $\pi = (0.2, 0.75, 0.05)$ where most weight is given to the state representing tiles with unchanged copy number. Based on that, we choose an initial transition matrix $A$ with equilibrium distribution $\pi$. That is, we set the self-transition probability of state $i \in S$ to $a_{ii} = 1 - s/\pi_i$ with respect to the scale parameter $s = 0.025$ to control the state durations, and we use $a_{ij} = (1 - a_{ii})/2$ for a transition from state $i$ to state $j \in S \setminus \{i\}$. We characterize the states by proper means and standard deviations using initial emission parameters $\mu_- = -2.5$, $\mu_= = 0$, $\mu_+ = 1.5$, $\sigma_- = 1$, $\sigma_= = 1$, and $\sigma_+ = 0.5$. We refer to the initial HMM by $\lambda^1$.

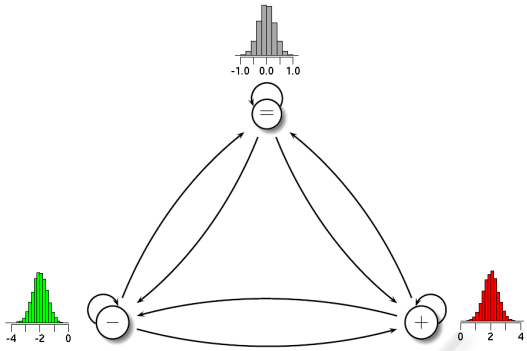


Figure 2: Three-state HMM with Gaussian emission densities for the analysis of Array-CGH data. States of the HMM are represented by circles labeled with $-$ (decreased), $=$ (unchanged), and $+$ (increased) modeling copy numbers of DNA segments in an ecotype compared to a reference genome. Transitions between states are represented by arrows modeling all possible transitions in an Array-CGH profile. Gaussian emission densities characterize the states. Thus, the emission density of the unchanged state (gray) has its mean about zero, whereas the emission densities of the decreased state (green) and the increased state (red) have means significantly different from zero.

### 2.2.3 HMM Prior

The incorporation of *a priori* knowledge is of prime importance for training an HMM to predict biologically relevant segments that vary in their copy numbers between ecotypes. This can be realized using a prior distribution. We define the prior $P[\lambda|\Phi]$ of the HMM $\lambda$ as a product of independent priors for each type of HMM parameters by

$$P[\lambda|\Phi] = P[\pi|\Phi] \cdot P[A|\Phi] \cdot P[E|\Phi].$$

We use a conjugate Dirichlet prior $P[\pi|\Phi]$ for start distribution $\pi$ defined by

$$P[\pi|\Phi] = c_\pi \prod_{i \in S} \pi_i^{\vartheta_\pi - 1}$$

with positive hyper-parameter $\vartheta_\pi \in \Phi$ and normalization constant $c_\pi$. The product of conjugate Dirichlet priors $P[A|\Phi]$ for transition matrix $A$ is given by

$$P[A|\Phi] = c_A \prod_{i \in S} \prod_{j \in S} a_{ij}^{\vartheta_a - 1}$$

with positive hyper-parameter $\vartheta_a \in \Phi$ and normalization constant $c_A$. We realize the prior for emission parameters $E$ using a product of conjugate Normal-Gamma priors

$$P[E|\Phi] = \prod_{i \in S} P[\mu_i|\Phi] \cdot P[\sigma_i|\Phi]$$

consisting of a state specific Gaussian density

$$P[\mu_i|\Phi] = \frac{\sqrt{\varepsilon_i}}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{\varepsilon_i}{2}\frac{(\mu_i-\eta_i)^2}{\sigma_i^2}\right)$$

as prior for mean $\mu_i$ with positive hyper-parameters $\eta_i \in \Phi$ (*a priori* mean) and $\varepsilon_i \in \Phi$ (scale of *a priori* mean), and a state-specific Inverted-Gamma prior

$$P[\sigma_i|\Phi] = \frac{2\alpha_i^{r_i}}{\Gamma(r_i)\sigma_i^{2r_i+1}} \exp\left(-\frac{\alpha_i}{\sigma_i^2}\right)$$

as prior for the standard deviation $\sigma_i$ with positive hyper-parameters $\alpha_i \in \Phi$ (scale of standard deviation) and $r_i \in \Phi$ (shape of standard deviation). The choice of this prior allows to include biological *a priori* knowledge into the training of the HMM. As motivated in Fig. 1, the integration of information about the emission parameters of the HMM is important. Additionally, the used prior allows to determine analytical parameter re-estimators for the HMM training. Especially, with the choice of the Normal-Gamma distribution as prior for a Gaussian emission density, we follow (Richardson and Green, 1997) transforming the proposed Gamma distribution as prior for the precision $\sigma_i^{-2}$ into an Inverted-Gamma prior for the standard deviation $\sigma_i$. With respect to the underlying biological question and motivated by the histogram of log-ratios in Fig. 1, we set the parameters of the Normal-Gamma priors to $\eta_- = -2.5$, $\eta_= = 0$, and $\eta_+ = 1.5$ (*a priori* means). We use $\varepsilon_- = 10,000$, $\varepsilon_= = 1,000$ and $\varepsilon_+ = 7,500$ (scale of *a priori* means), $r_- = 20,000$, $r_= = 1$, and $r_+ = 1,000$ (shape of standard deviations), and $\alpha_- = \alpha_= = \alpha_+ = 10^{-4}$ (scale of standard deviations), but in general this depends on the number of tiles in an Array-CGH experiment. We ensure that the HMM can start in each state and that all transitions are allowed by setting $\vartheta_\pi = 10/3$ and $\vartheta_a = 10/9$. The choice of these prior parameters ensures a good characterization of the three HMM states.

### 2.2.4 HMM Training

In most cases HMMs are trained by iteratively maximizing the likelihood of the observation sequences under the model using the Baum-Welch algorithm (Baum, 1972; Rabiner, 1989; Durbin et al., 1998). This algorithm is part of the class of EM algorithms, which can be extended to include *a priori* knowledge into the iterative training by maximizing the posterior (Dempster et al., 1977). We train the initial HMM on all Array-CGH profiles using a *maximum a posteriori* (MAP) variant of the standard Baum-Welch algorithm. That is, we iteratively obtain new HMM parameters

$$\lambda^{h+1} = \underset{\lambda}{\operatorname{argmax}} \left(Q(\lambda|\lambda^h) + \log\left(P[\lambda|\Phi]\right)\right)$$

that maximize the posterior based on given Array-CGH profiles $O = (o^1, \ldots, o^K)$ and current HMM parameters $\lambda^h$. Here, $Q(\lambda|\lambda^h)$ represents Baum's auxiliary function (Rabiner, 1989; Durbin et al., 1998)

$$Q(\lambda|\lambda^h) = \sum_{k=1}^{K} \sum_{q \in S^{T_k}} P[q|o^k, \lambda^h] \log\left(P[o^k, q|\lambda]\right)$$

with complete data likelihood

$$P[o, q|\lambda] = \pi_{q_t} \prod_{t=1}^{T-1} a_{q_t q_{t+1}} \prod_{t=1}^{T} b_{q_t}(o_t)$$

of Array-CGH profile $o$ and corresponding state sequence $q$. The conflation of $Q(\lambda|\lambda^h)$ and $P[\lambda|\Phi]$ enables us to include state-specific *a priori* knowledge about the parameters of Gaussian emission densities into the training. Based on that, we use Lagrange multipliers to determine the re-estimation formula for each type of HMM parameters given by

$$\pi_i^{h+1} = \frac{\vartheta_\pi - 1 + \sum_{k=1}^{K} \gamma_1^k(i)}{|S|\vartheta_\pi - |S| + K}$$

$$a_{ij}^{h+1} = \frac{\vartheta_a - 1 + \sum_{k=1}^{K} \sum_{t=1}^{T^k-1} \varepsilon_t^k(i,j)}{|S|\vartheta_a - |S| + \sum_{k=1}^{K} \sum_{t=1}^{T^k-1} \gamma_t^k(i)}$$

$$\mu_i^{h+1} = \frac{\varepsilon_i\eta_i + \sum_{k=1}^{K} \sum_{t=1}^{T^k} \gamma_t^k(i) \cdot o_t^k}{\varepsilon_i + \sum_{k=1}^{K} \sum_{t=1}^{T^k} \gamma_t^k(i)}$$

$$\sigma_i^{h+1} = \sqrt{\frac{\Delta_i + \sum_{k=1}^{K} \sum_{t=1}^{T^k} \gamma_t^k(i) \cdot (o_t^k - \mu_i^{h+1})^2}{2r_i + 2 + \sum_{k=1}^{K} \sum_{t=1}^{T^k} \gamma_t^k(i)}}$$

with $\Delta_i = \varepsilon_i(\mu_i^{h+1} - \eta_i)^2 + 2\alpha_i$. We calculate the probabilities $\gamma_t^k(i) = P[q_t = i|o^k, \lambda^h]$ and $\varepsilon_t^k(i,j) = P[q_t = i, q_{t+1} = j|o^k, \lambda^h]$ using the Forward-Backward algorithms for HMMs (Rabiner, 1989; Durbin et al., 1998). Starting with the initial HMM $\lambda^1$ ($h = 1$), we iteratively determine new HMM parameters $\lambda^{h+1}$ and stop if the increase of the log-posterior of two successive training iterations is less than $10^{-3}$.

### 2.2.5 Segment Detection and Scoring

After the training of the initial HMM on all Array-CGH profiles, we apply the Viterbi algorithm (Rabiner, 1989; Durbin et al., 1998) to determine the most probable state sequence $q$ for each Array-CGH profile $o$. The so-called Viterbi path $q$ partitions

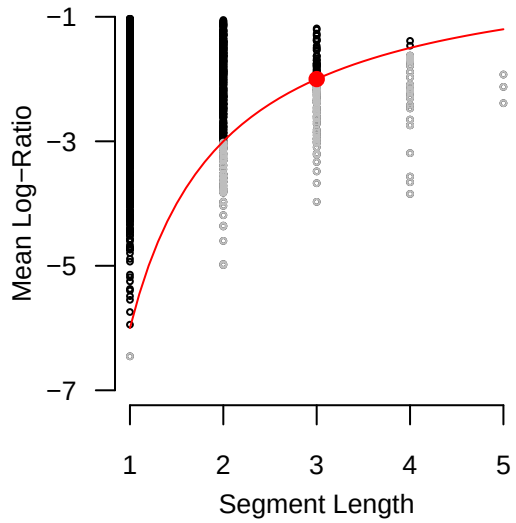


Figure 3: Visualization of the scoring scheme for DNA segments with decreased copy numbers. The red dot characterizes a DNA segment with segment length $N = 3$, mean log-ratio $L = -2$, and score $S = N \cdot L = -6$, which was predicted to have a decreased copy number status by the trained HMM in the original Array-CGH data. The red curve represents the hyperbola $f(n) = S/n$ of segment length $n$ that divides the segments with decreased copy number status obtained from the permuted Array-CGH profiles into segments with scores greater than $-6$, represented by black dots, and segments with scores less or equal than $-6$, represented by gray dots. The *Score*-value of the segment represented by the red dot is the proportion of gray dots in relation to the total number of gray and black dots. Here, the *Score*-value is 0.088.

the corresponding Array-CGH profile into DNA segments of decreased, unchanged, or increased copy numbers in relation to the reference genome sequence. We refer to such a segment of copy number status $i \in S$ by

$$q_{t_s}^{t_e}(i) = q_{t_s}, \ldots, q_{t_e}$$

where the length of this segment is maximal ($q_{t_s-1} \neq i$ and $q_{t_e+1} \neq i$ for $t_s \leq t_e$), and all tiles within this segment have the identical copy number status $i$ ($q_t = i$ for each $t \in [t_s, t_e]$). We score each segment by the sum of its log-ratios

$$S(q_{t_s}^{t_e}(i)|o) = N_{t_s}^{t_e} \cdot L_{t_s}^{t_e} = \sum_{t=t_s}^{t_e} o_t$$

to incorporate the segment length $N_{t_s}^{t_e} = t_e - t_s + 1$ with respect to the mean log-ratio $L_{t_s}^{t_e} = (1/N_{t_s}^{t_e}) \sum_{t=t_s}^{t_e} o_t$ within the segment. Next, we determine the relevance of predicted DNA segments with decreased or increased copy numbers. That is, we permute the log-ratios of each Array-CGH profile, and then, we apply the trained HMM to this data to predict DNA segments with changed copy numbers that we score by $S(q_{t_s}^{t_e}(i)|o)$. We repeat this step 100 times resulting in two score distributions: one for DNA segments with decreased copy numbers and another one for DNA segments with increased copy numbers. The *Score*-value of a predicted segment in the original data is calculated by determining the proportion of scores under the corresponding score distribution that are identical or more extreme than the score of the considered segment. DNA segments with a *Score*-value close to zero are of most interest for our biologists. Fig. 3 shows an illustration of the scoring scheme.

## 3 RESULTS AND DISCUSSION

With the aim of predicting DNA copy number variations between *Arabidopsis thaliana* ecotype C24 and Columbia, we separately trained one HMM for each of the two arrays of the Array-CGH experiment. Then, we used each trained HMM to determine DNA segments with decreased or increased copy numbers that we scored under permuted data for obtaining segments of decreased or increased copy numbers at a *Score*-value level of 0.01. Alternatively, the standard service of Roche NimbleGen, Inc. included the segmentation of Array-CGH profiles using their segMNT algorithm (Roche NimbleGen, Inc., 2008) resulting in two genome-wide segmentation profiles. The first goal of the following study is to investigate where the predictions of the HMM approach overlap and where they differ from those of the segMNT algorithm. The second goal is to characterize the prediction behavior of both methods for various levels of log-ratio signal intensities.

### 3.1 Comparison of Segmentations

Based on the SignalMap viewer (Roche NimbleGen, Inc.), we performed a genome-wide inspection of the segmentation results obtained by the HMM approach and by the segMNT algorithm. In practice, this basic step allows a first direct comparison of both methods, and it provides a general overview of DNA regions where copy number variations have occurred. Both approaches predicted DNA segments with decreased or increased copy numbers widely spread over the reference genome of *Arabidopsis thaliana* ecotype Columbia. For the HMM approach we could quantify the proportion of these segments directly, because each of these segments is assigned to one of the three HMM states modeling the underlying copy number (Fig. 2). Considering the segMNT algorithm, this

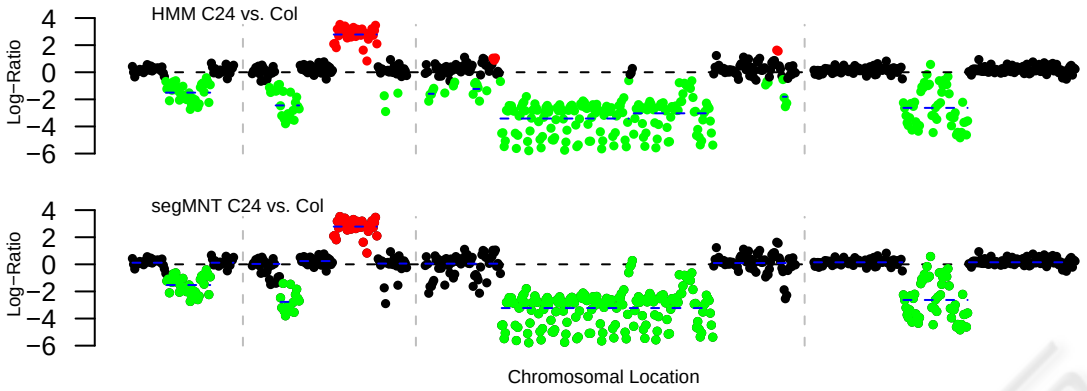


Figure 4: Exemplary comparison of segmentation results for genome-wide selected DNA regions for ecotype C24 compared to Columbia. From left to right separated by gray dashed lines: Region 1 [Chr1 2,779,800 bp - 2,806,847 bp], Region 2 [Chr2 2,067,828 bp - 2,109,088 bp], Region 3 [Chr4 11,026,348 bp - 11,125,315 bp], and Region 4 [Chr4 13,596,032 bp - 13,665,842 bp]. The top plot represents the segmentation results of the HMM approach. Green dots label tiles predicted by the HMM to have a decreased copy number. Red dots label tiles predicted by the HMM to have an increased copy number. Blue dashed lines highlight DNA segments significantly different from permuted data at a *Score*-value threshold of 0.01. Black dots label tiles predicted by the HMM to have unchanged copy numbers. The bottom plot represents the segmentation results of the segMNT algorithm colored like the HMM segmentation. Both approaches provide a nearly identical segmentation of the displayed DNA regions.

was not directly possible, because the obtained segments are not grouped by copy number status. That is, the segMNT algorithm is only performing a segmentation of the Array-CGH profiles, but the predicted segments are not categorized into segments with decreased, unchanged, and increased copy numbers. Thus, we categorized all segments predicted by the segMNT algorithm with a mean log-ratio less or equal than $-0.5$ as segments with decreased copy numbers, and in analogy, we labeled all segments with mean log-ratios greater or equal than 0.5 as segments with increased copy numbers. For a first comparison of segments that have been predicted by both methods, we only count a segment predicted by the HMM at a *Score*-value level of 0.01 as a segment with decreased copy number if its corresponding mean log-ratio is less or equal than $-0.5$, and otherwise as a segment with increased copy number if its mean log-ratio is greater or equal than 0.5. The numbers of predicted segments for each approach are shown in Tab. 1. DNA segments which are deleted in Columbia, but which exist in C24 cannot be measured, because the tiling arrays represent the reference genome of ecotype Columbia. Thus, it is expected that the small proportion of DNA segments with increased copy numbers is caused by this kind of array design (Fig. 1). The HMM approach identified significantly more DNA segments with decreased or increased copy numbers than the segMNT algorithm. These numbers alone cannot be used to decide which method should be preferred for the comparison of *Arabidopsis thaliana* ecotypes as also the lengths

Table 1: Number of segments with decreased ($-$) and increased ($+$) copy numbers in ecotype C24 predicted by HMM and segMNT.

| **Method** | **Array 1** | | **Array 2** | |
|---|---|---|---|---|
| | $-$ | $+$ | $-$ | $+$ |
| HMM | 1352 | 196 | 1427 | 205 |
| segMNT | 271 | 3 | 247 | 4 |

of predicted segments, their chromosomal locations, and their reproducibility between both arrays should be considered. To address this, we first analyzed the overlap of segments predicted for both arrays by HMM or segMNT. The results are shown in Tab. 2. In general, both methods showed a good reproducibility of their predicted segments. The higher number of predicted segments by the HMM (Tab. 1) did not lead to a great loss of reproducibility. That is, the HMM predicted more reproducible segments for both arrays than the segMNT algorithm. Next, we investigated how many of the segments that were predicted by the segMNT algorithm have also been identified by the HMM. All segMNT segments of Tab. 1 were also predicted by the HMM with respect to slightly varying segment start and end points. The general overlap between both approaches is exemplarily shown for selected DNA segments in Fig. 4. Additionally, Fig. 5 shows representative DNA regions where the HMM identified a copy number change reproducible for both interleaved arrays whereas the segMNT algorithm failed.

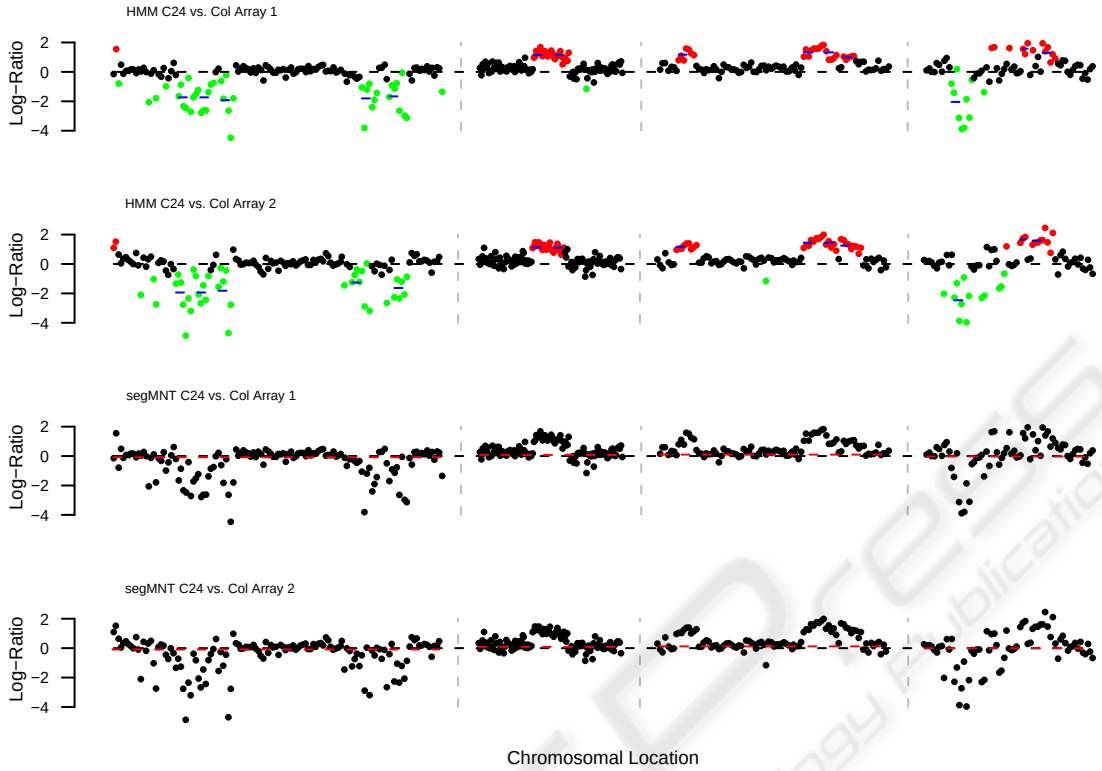


Figure 5: Exemplary comparison of segmentation results for DNA regions on chromosome 4 for ecotype C24 compared to Columbia. From left to right separated by gray dashed lines: Region 1 [654,108 bp - 697,518 bp], Region 2 [1,305,320 bp - 1,324,132 bp], Region 3 [3,731,013 bp - 3,761,229 bp], and Region 4 [5,411,025 bp - 5,433,126 bp]. The two top plots represent segmentation results of the HMM approach for both interleaved arrays. Green dots label tiles predicted by the HMM to have a decreased copy number. Red dots label tiles predicted by the HMM to have an increased copy number. Blue dashed lines highlight DNA segments significantly different from permuted data at a *Score*-value threshold of 0.01. Black dots label tiles predicted by the HMM to have unchanged copy numbers. The two bottom plots represent segmentation results of the segMNT algorithm for both arrays. Red dashed lines show that no segmentation was obtained. Both approaches provide a quite different segmentation of the DNA regions. Here, the segMNT algorithm failed to identify segments with decreased or increased copy numbers. The HMM approach clearly identifies segments with significantly decreased or increased copy numbers, and in addition, these biologically interesting results are reproducible for both arrays.

Table 2: Number of reproducible segments between Array 1 and Array 2 for HMM and segMNT based on predicted segments of Tab. 1. A$i$ vs. A$j$: Segments predicted for Array $i$ that are also predicted for Array $j$. Differences in counts result mostly from segments predicted in one array that are predicted in the other array by several non-overlapping segments.

| | **A1 vs. A2** | | **A2 vs. A1** | |
|---|---|---|---|---|
| **Method** | $-$ | $+$ | $-$ | $+$ |
| HMM | 1114 | 119 | 1209 | 129 |
| segMNT | 232 | 3 | 224 | 3 |

## 3.2 Genome-wide Performance

The true copy number status of a predicted segment can be experimentally validated using independent methods like PCR, sequencing, or insitu hybridization in the wet-lab. For thousands of identified segments with putative copy number variations the usage of such validation methods is currently not practical. In order to investigate how the prediction results of the segMNT algorithm and the HMM approach differed at various log-ratio levels, we used a biologically motivated model that defined the copy number status of each tile with respect to its measured log-ratio. Based on the measurements of Array-CGH experiments, segments with decreased copy numbers consist of tiles with log-ratios much less than zero, and segments with increased copy number are represented by tiles with log-ratios much greater than zero (Fig. 1). For these reasons, we use a variable log-ratio threshold $\Delta \in \mathbb{R}^+$ to define the copy number status for each tile in an Array-CGH profile. That is, a tile with log-ratio $o_t$ is defined to have a decreased copy number if its log-ratio $o_t$ is less or equal to the log-ratio threshold $-\Delta$, or conversely this tile is defined

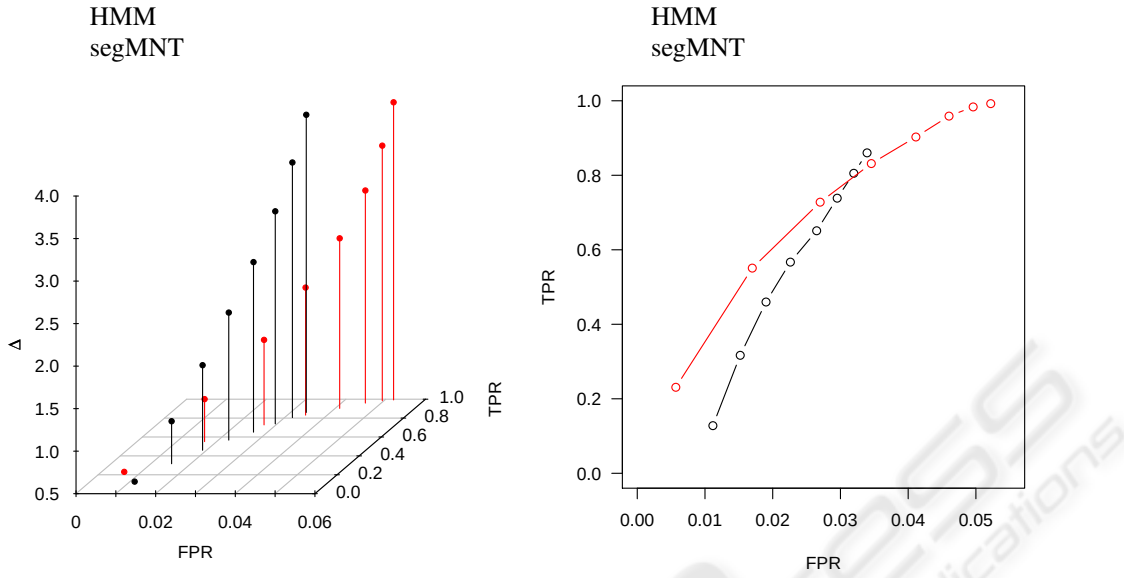


Figure 6: Prediction results obtained by segMNT and HMM at various log-ratio levels $\Delta \in [0.5, 4]$. Left: FPR vs. TPR in the context of the variable log-ratio threshold $\Delta$ defining the copy number of segments as being decreased or increased. Right: Top view on the left figure neglecting the log-ratio threshold.

to have an increased copy number if its log-ratio $o_t$ is greater or equal to the log-ratio threshold $\Delta$. All tiles that are not defined to have decreased or increased copy numbers are defined to have an unchanged copy number. Based on that, we evaluated both segmentations, the one of the HMM at a *Score*-value level of 0.01 and the other one of the segMNT algorithm from the previous section, against the segmentations obtained from systematically chosen variable log-ratio thresholds $\Delta \in [0.5, 4.0]$. For each log-ratio threshold $\Delta$ we determined the True-Positive-Rate $\mathrm{TPR} = \mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$ and the $\mathrm{FPR} = \mathrm{FP}/(\mathrm{FP}+\mathrm{TN})$ of both approaches. The results are shown in Fig. 6. In general, the HMM showed a much higher TPR than the segMNT algorithm at a moderately higher FPR. This confirms the previous findings: the HMM approach identifies much more DNA segments with copy number variations.

## 4 CONCLUSIONS

We introduced a three-state HMM approach (Fig. 2) for comparing the genomes of different ecotypes of *Arabidopsis thaliana*. This approach is capable of (i) incorporating biologically *a priori* knowledge into the training of model parameters, and of (ii) scoring DNA segments of decreased or increased copy numbers separately using permuted Array-CGH data. We observed that our HMM approach identifies significantly more reproducible DNA segments with decreased or increased copy numbers than the routinely used segMNT algorithm. Using this HMM approach, we find that about 5% of the genome of ecotype C24 shows decreased copy numbers and 0.3% shows increased copy numbers compared to the reference genome of ecotype Columbia. Thus, we obtained a detailed map characterizing regions of DNA copy number variations for future studies of ecotypes including the analysis of DNA-histone interactions, histone modifications, and transcript profiling. Further biological interpretation of such a map can be obtained using the AtEnsEMBL genome browser (James et al., 2007) for representing the map in the context of the genome annotation. In summary, all results indicate that our HMM approach provides a good basis for Array-CGH-based genome comparison of *Arabidopsis thaliana* ecotypes. One of our future analyses will focus on an in-depth comparison of the HMM approach against other available methods for analyzing Array-CGH data.

## ACKNOWLEDGEMENTS

This work was supported by the Ministry of culture Saxony-Anhalt grant XP3624HP/0606T and by the DFG grant HO1779/7-2.

# REFERENCES


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

Borevitz, J. O., Liang, D., Plouffe, D., Chang, H.-S., Zhu, T., Weigel, D., Berry, C. C., Winzeler, E., and Chory, J. (2003). Large-scale identification of single-feature polymorphisms in complex genomes. *Genome Res*, 13:513–523.

Cahan, P., Godfrey, L. E., Eis, P. S., Richmond, T. A., Selzer, R. R., Brent, M., McLeod, H. L., Ley, T. J., and Graubert, T. A. (2008). wuHMM: a robust algorithm to detect DNA copy number variation using long oligonucleotide microarray data. *Nucleic Acids Research*, 36(7):1–11.

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

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

Fan, C., Vibranovski, M. D., Chen, Y., and Long, M. (2007). A Microarray Based Genomic Hybridization Method for Identification of New Genes in Plants: Case Analyses of Arabidopsis and Oryza. *J Integr Plant Biol*, 49(6):915–926.

Fridlyand, J., Snijders, A. M., Pinkel, D., Albertson, D. G., and Jain, A. N. (2004). Hidden Markov models approach to the analysis of array CGH data. *J Multivariate Analysis*, 90:132–153.

Hupé, P., Stransky, N., Thiery, J.-P., Radvanyi, F., and Barillot, E. (2004). Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics*, 20(18):3413–3422.

James, N., Graham, N., Celments, D., Schildknecht, B., and May, S. (2007). AtEnsEMBL: A Post-Genomic Resource Browser for Arabidopsis. *Methods Mol Biol*, 406:213–228.

Jong, K., Marchiori, E., Meijer, G., Vaar, A. v. d., and Ylstra, B. (2004). Breakpoint identification and smoothing of array comparative genomic hybridization data. *Bioinformatics*, 20(18):3636–3637.

Lai, W. R., Johnson, M. D., Kucherlapati, R., and Park, P. J. (2005). Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics*, 21(19):3763–3770.

Mantripragada, K. K., Buckley, P. G., de Stahl, T. D., and Dumanski, J. P. (2004). Genomic microarrays in the spotlight. *Trends Genet*, 20:87–94.

Marioni, J. C., Thorne, N. P., and Tavaré (2006). BioHMM: a heterogeneous hidden Markov model for segmenting array CGH data. *Bioinformatics*, 22(9):1144–1146.

Martienessen, R. A., Doerge, R. W., and Colot, V. (2005). Epigenomic mapping in Arabidopsis using tiling microarrays. *Chromosome Research*, 13:299–308.

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

Richardson, S. and Green, P. J. (1997). On Bayesian Analysis of Mixtures with an Unknown Number of Components. *Journal of the Royal Statistical Society, Series B*, 59(4):731–792.

Roche NimbleGen, Inc. (2008). A Performance Comparison of Two CGH Segmentation Analysis Algorithms: DNACopy and segMNT. Available online: http://www.nimblegen.com.

Rueda, O. M. and Díaz-Uriate, R. (2007). Flexible and Accurate Dection of Genomic Copy-Number Changes from aCGH. *PLoS Comput Biol*, 3(6).

Willenbrock, H. and Fridlyand, J. (2005). A comparison study: applying segmentation to array CGH data for downstream analyses. *Bioinformatics*, 21(22):4084–4091.